The Evolution of GII.4 Norovirus and the Sources and Drivers of Norovirus Pandemics

Christopher Ruis

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Division of Infection and Immunity

University College London

January 8, 2018

I, Christopher Ruis, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

0.1 Abstract

The norovirus genotype GII.4 is a leading cause of human gastroenteritis worldwide and has caused six pandemics since the mid-1990s. In this thesis, we use phylogenetic analyses to investigate the evolutionary history of the GII.4 genotype and the sources and drivers of norovirus pandemics.

We first examine the early history of GII.4 and suggest that the increased prevalence of GII.4 concomitant with the first pandemic was a 'perfect storm' where a virus capable of accommodating a high level of amino acid change with a high mutation rate enabling efficient transmission acquired a highly stable viral capsid and/or an increased susceptible population size by expanding its receptor-binding repertoire.

We next reconstruct the temporal history of GII.4 and demonstrate that each pandemic strain circulated undetected within poorly sampled reservoir populations for years prior to pandemic emergence. Over several years prior to pandemic emergence, the strain diversifies into a large number of lineages and spatiotemporal reconstruction suggests the strain undergoes low-level worldwide circulation. This indicates that the viral genetic changes important for pandemic emergence are acquired years prior to the pandemic and are therefore not the proximal driver of pandemic spread; we hypothesise that genetic changes pre-adapt the strain for future emergence by shifting the virus to a new region of antigenic space. We demonstrate significant amino acid diversity within pandemic strains, with highly diverse sites within a strain often coinciding with immune epitopes and/or receptor-binding regions. This diversity begins to be accumulated prior to pandemic emergence. We hypothesise that increasing host population immunity curtails circulation of the preceding pandemic strain and results in a new pandemic by opening a niche into which many closely related but subtly different viral lineages can emerge.

Finally, we examine two newly emerging norovirus strains and demonstrate that they share polymerase substitutions that may enable increased transmission.

0.2 Acknowledgements

I would first and foremost like to extend my sincerest thanks to my supervisors, Richard Goldstein and Judy Breuer, for their support, guidance and expertise throughout my PhD and for furthering my interest in pathogens and molecular evolution. I have greatly enjoyed my time working with them both.

Thanks also to all of the members of the Goldstein and Breuer labs for their numerous useful and enjoyable discussions over the past three years.

This work was funded by a studentship from UCL CoMPLEX.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Norovirus gastroenteritis and global prevalence

1.1.1 Norovirus prevalence and disease burden

The report of Zahorsky in 1929 described the 'winter vomiting disease', an illness characterised by the rapid onset of self-limiting vomiting and diarrheoa that exhibited peaks in winter months (Zahorsky, 1929). It was not until 1972 that Kapikian et al identified the aetiological agent likely responsible for winter vomiting disease in a gastroenteritis outbreak in an elementary school in Norwalk, Ohio (Kapikian et al., 1972). This Norwalk agent was the first causative agent of acute infectious nonbacterial gastroenteritis to be identified (Kapikian et al., 1972) and is the prototype virus for the *Norovirus* genus, one of five genera within the *Caliciviridae*. While the disease burden of norovirus was initially hard to assess due to a lack of simple diagnostic tests (Widdowson et al., 2005), the cloning of the norovirus genome in 1990 and the advent of reverse transcription-polymerase chain reaction (RT-PCR) enabled assessment of the contribution of norovirus to the global burden of gastroenteritis (Xi et al., 1990; Widdowson et al., 2005; Lopman et al., 2016). Norovirus is now estimated to cause roughly 684 million episodes of diarrheoal disease annually, afflicting close to 1 in 10 of the world's population each year (Kirk et al., 2015; Pires et al., 2015; Lopman et al., 2016). Norovirus is ubiquitous and is associated with gastroenteritis in low-, middle- and high-income countries and within community, outpatient and inpatient settings (Ahmed et al., 2014; Lopman et al., 2016;

Mans et al., 2016). Prevalence estimates suggest that norovirus is the leading cause of diarrhoeal cases in patients of all age groups (Pires et al., 2015; Lopman et al., 2016), causing roughly 18% of acute gastroenteritis cases and 50% of gastroenteritis outbreaks worldwide (Patel et al., 2009; Ahmed et al., 2014). Seroprevalence studies suggest that almost all adults have been exposed to at least one norovirus (Donaldson et al., 2010).

The majority of norovirus sampling is carried out in outbreak environments, typically in hospitals and aged care facilities, and little is known about norovirus prevalence and epidemiology in outbreaks and sporadic gastroenteritis in the community (Inns et al., 2017). The prevalence of norovirus in the community is therefore thought to be substantially underestimated (Inns et al., 2017). The Infectious Intestinal Disease studies (IID1 and IID2) were carried out in the UK from 1993-1996 and 2008-2009 and estimated the community prevalence of norovirus (Wheeler et al., 1999; Tam et al., 2012). The IID2 study estimated that roughly 3 million clinically significant norovirus cases (defined as detection of norovirus with a cycle threshold (ct) value of 30 of less) occur in the UK annually (Tam et al., 2012). More recently, Harris et al. (2017) used a less stringent, but diagnostically relevant, ct cutoff of 40 to define norovirus cases and re-estimated norovirus prevalence in the UK at 3.7 million (95% confidence interval (CI) 3.3-4.1 million) annual infections. While norovirus is associated with gastroenteritis in all age groups, age-specific incidence rates from the IID2 study demonstrated that the infection rate is significantly higher in children under five years of age compared with other age groups in both the community (142.6 cases per 1000 person years (95% CI 99.8-203.9) versus 37.6 (95% CI 31.5-44.7)) and presenting to primary healthcare (14.4 cases per 1000 person years (95% CI 8.5-24.5) versus 1.4 (95% CI 0.9-2.0)).

There has been significant progress in reducing global diarrhoeal deaths, with a 50% reduction from 2.6 million deaths in 1990 to 1.3 million deaths in 2013 (GBD 2013 Mortality and Causes of Death Collaborators, 2015). However, diarrhoeal deaths remain significant and still represent the fourth most common cause of mortality (GBD 2013 Mortality and Causes of Death Collaborators, 2015). Norovirus is estimated to cause roughly 212,000 deaths annually, with the vast majority of these deaths occurring in countries defined as middle- or high-mortality by the World Health Organisation (Pires et al., 2015). Most deaths occur in children under five years of age and norovirus is the second most frequent cause of diarrheoal death in this age group, in addition to being the most fre-

quent cause of diarrheoal death in children over the age of five (Lopman et al., 2016). While the number of deaths in developed countries is far lower, norovirus has resulted in fatalities within young children, the immunocompromised and elderly patients. Annually, norovirus is estimated to result in 64,000 episodes of diarrheoa requiring hospitalisation and 900,000 clinic visits in children from developed countries (Patel et al., 2008). Additionally, with the introduction of rotavirus vaccination, noroviruses are now the most frequent cause of diarrheoal illness requiring medical attention in children under five years of age in the USA (Payne et al., 2013). Recent estimates suggest norovirus has a global economic cost of US$60.3 billion in societal costs and US$4.2 billion in direct health system costs annually, with most of this economic burden due to disease in children under five years of age (Bartsch et al., 2016).

1.1.2 Symptoms of norovirus gastroenteritis

Norovirus infection has a typical incubation period of 24-48 hours (median 1.2 days) (Patel et al., 2009; Lee et al., 2013). Symptoms can vary from individual to individual, but often consist of vomiting and/or non-bloody diarrheoa, which may be accompanied by nausea, stomach cramps and muscle pain (Lopman et al., 2004b; Patel et al., 2009; Robilotti et al., 2015). The symptomatic period typically lasts 12-72 hours, but there is evidence of a longer symptomatic period in hospitalised patients and young children (Lopman et al., 2004b; Patel et al., 2009; Robilotti et al., 2015). In most individuals, disease is self-limiting and patients recover fully after the symptomatic period. However, more prolonged and severe disease has been noted in immunocompromised patients, who can be infected persistently for up to several years, and occasionally in the elderly and other otherwise healthy individuals (Bok and Green, 2012; Robilotti et al., 2015).

While norovirus is highly prevalent in symptomatic gastroenteritis cases, it has become clear that large numbers of individuals are asymptomatically infected (Teunis et al., 2015; Ahmed et al., 2014). Indeed, in some studies in low income countries the proportion of norovirus-positive patients is similar in symptomatic and asymptomatic groups (Kotloff et al., 2013). In a systematic review of norovirus prevalence, Ahmed et al estimated a total prevalence of 7% in asymptomatic control patients compared with 20% in

symptomatic patients in the same studies (Ahmed et al., 2014), while volunteer studies suggest that up to 50% of exposed individuals will not exhibit any symptoms (Patel et al., 2009). Additionally, patients can shed virus in stool for up to several months after the resolution of symptoms, with similar periods of shedding being estimated for symptomatic and asymptomatic individuals (Teunis et al., 2015).

1.2 Norovirus molecular and cellular biology

1.2.1 Genome organisation

The *Caliciviridae* are a family of small nonenveloped positive sense single stranded RNA viruses (Green, 2007). The family consists of five accepted genera: *Norovirus*, *Sapovirus*, *Lagovirus*, *Nebovirus* and *Vesivirus*, of which *Norovirus* and *Sapovirus* are known to infect humans. There are also three proposed genera that are yet to be approved: *Recovirus*, *Valovirus* and chicken calicivirus (Smits et al., 2012). The linear nonsegmented norovirus genome is 7.3-7.7 kilobases (Kb) in length and is bound by a virally encoded protein VPg (Viral Protein genome-linked) at the 5' end and is polyadenylated at the 3' end (Green, 2007; Thorne and Goodfellow, 2014) (Figure 1.1). The 5' and 3' untranslated regions (UTRs) are short, with the 5' UTR consisting of just four nucleotides in human noroviruses and five nucleotides in murine noroviruses (MNVs) (Thorne and Goodfellow, 2014) (Figure 1.1). The norovirus genome encodes three open reading frames (ORFs), with the exception of MNVs which encode an additional fourth ORF (Figure 1.1). ORF1 is more than 5Kb in length and encodes an approximately 200 kiloDalton nonstructural polyprotein that is co- and post-translationally cleaved by a virally-encoded protease into six proteins (Belliot et al., 2003; Sosnovtsev et al., 2006), which are typically given different names for human noroviruses and MNVs: p48 (NS1/2 in MNV), NTPase (NS3), p22 (NS4), VPg (NS5), protease (NS6) and RNA-dependent RNA polymerase (RdRp, NS7) (Figure 1.1). ORF2 is roughly 1.8Kb in length, overlaps with ORF1 by roughly 20 nucleotides and encodes the major capsid protein VP1, 180 copies of which form the viral particle (Prasad et al., 1999; Green, 2007). The cap-

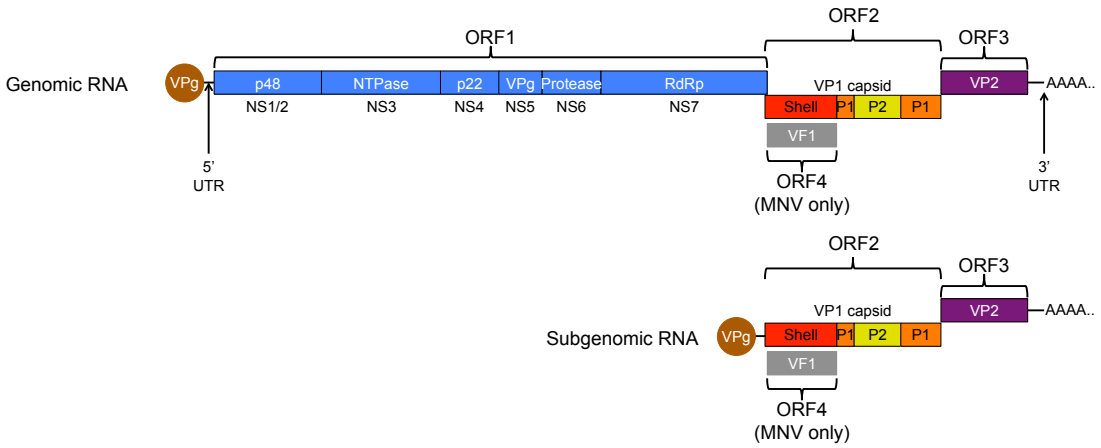


Figure 1.1: Norovirus genome organisation. A schematic of the norovirus genome is shown depicting the open reading frames (ORFs). The complete genome is between 7.3 and 7.7 kilobases in length depending on genogroup. ORF1 encodes a nonstructural polyprotein that is co- and post-translationally cleaved by the virally encoded protease into six proteins: p48, NTPase, p22, VPg, protease and RdRp. The NS name shown below each protein is the name of the protein in MNV. ORF2 encodes the VP1 capsid protein that is subdivided into the shell domain and two protruding subdomains P1 and P2. The P2 subdomain (shown in yellow) is an insert into the P1 subdomain (shown in orange), with the P1 subdomain being encoded on either side of the P2 subdomain. ORF3 encodes the minor structural protein VP2. ORF4 is present only in MNVs and encodes virulence factor 1 (VF1), an antagonist of the innate immune response. The genome has a very short 5' UTR (sequence GTGA in GII viruses) and a short 3' UTR. The 5' end of the genome is bound by the viral protein VPg, which acts as a cap substitute for translation and the 3' end of the genome is poly-adenylated. ORFs 2 and 3 are produced as a subgenomic RNA that is identical to the corresponding region in the full length genome, is bound by VPg at the 5' end and is polyadenylated.

sid protein is divided into the shell and protruding domains, with the protruding domain being further subdivided into the P1 and P2 subdomains (Figures 1.1, 1.2). ORF3 is approximately 0.6Kb in length and encodes a small basic protein VP2, a small number of copies of which are incorporated into the viral particle (Vongpunsawad et al., 2013). The ORF3 start codon overlaps with the stop codon of ORF2 by a single nucleotide (Figure 1.1). The second and third ORFs are also produced as a subgenomic RNA that is identical to the corresponding region in the complete genome and is bound by VPg at the 5' end and polyadenylated at the 3' end (Figure 1.1) (Thorne and Goodfellow, 2014). The fourth ORF in MNVs encodes virulence factor 1 (VF1), an antagonist of the innate immune response that is translated from an alternative reading frame located within ORF2 (McFadden et al., 2011).

The norovirus particle is roughly 38nm in diameter and is formed from 180 copies

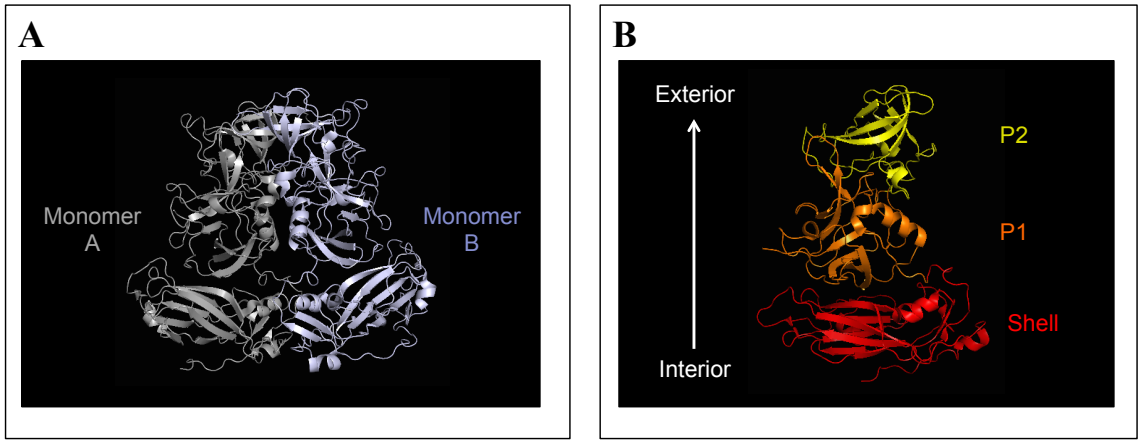


Figure 1.2: The structure of the norovirus capsid protein. (**A**) The capsid is formed from 90 dimers. The structure of the dimer is shown with one monomer in grey and the other monomer in blue-white. Interactions occur between the monomers in each subdomain of the capsid. (**B**) The capsid is divided into the shell and protruding domains, with the protruding domain being further subdivided into the P1 and P2 subdomains. The shell domain forms the interior of the viral particle and the P2 domain is surface exposed.

of the VP1 capsid protein, organised into 90 dimers with a T=3 icosahedral symmetry (Prasad et al., 1999). The particle also contains a small number of copies of VP2 (Vongpunsawad et al., 2013). The shell domain is found on the interior face of the viral particle (Figure 1.2). The shell domain forms the icosahedral shell and when expressed alone can form smooth icosahedral particles, while the P domain forms four regions of dimeric interactions that stabilise the particle (Prasad et al., 1999; Bertolotti-Ciarlet et al., 2002; Cao et al., 2007). The shell domain is formed from capsid residues 1-221 (site numbers relative to the GII.4 capsid) and is connected to the protruding domain by a flexible hinge approximately ten residues in length (Prasad et al., 1999). The protruding domain consists of residues 222-540 and multiple studies have demonstrated that the P2 subdomain (residues 275-417) is an insertion in the P1 subdomain (residues 222-274 and 418-540) (Prasad et al., 1999). The P2 subdomain is found on the surface of the viral particle (Figure 1.2). The shell domain exhibits a high level of conservation, while the P2 domain exhibits the highest level of variability within the norovirus genome (Bull et al., 2010).

1.2.2 Norovirus classification

Until recently, most noroviruses could not be cultured and therefore classification

has been based on seqeuence and phylogenetic analyses (Zheng et al., 2006; Kroneman et al., 2013). While classification was initially based on percentage amino acid similarity (Zheng et al., 2006), this method was recently shown to be unreliable and so phylogenetic clustering is now used (Kroneman et al., 2013). Recombination has been demonstrated to occur frequently at the ORF1-ORF2 boundary, enabling novel combinations of the nonstructural and structural proteins (Bull et al., 2005, 2007; Eden et al., 2013). This recombination likely occurs through a mechanism involving the full length and subgenomic RNAs, whereby the RdRp reaches a RNA stem loop structure at the start of ORF2, losses processivity, falls off the full length genome and reattaches at the stem loop structure of a different subgenomic RNA molecule (Bull et al., 2005). Due to this frequent recombination, a dual-typing nomenclature system has been proposed where the RdRp and capsid are classified independently into genogroups and further into genotypes (Figure 1.3) (Zheng et al., 2006; Kroneman et al., 2013). There are currently seven genogroups, termed GI-GVII (Figure 1.3) (Vinje, 2015). There are 9 recognised genotypes in GI, 23 in GII, 2 in GIII, 2 in GIV, 1 in GV, 2 in GVI and 1 in GVII (Figure 1.3) (Vinje, 2015). The genotypes GI.3, GI.5, GI.7, GII.3 and GII.6 are further subdivided into multiple subclusters (Vinje, 2015). Under the dual-typing system (Kroneman et al., 2013), a virus classified as GII.P1-GII.4 would have a RdRp in genotype 1 and a capsid in genotype 4, both within genogroup II. Human norovirus infections are predominantly caused by viruses within genogroups GI and GII, with a smaller number of infections with GIV.1 viruses (Vinje, 2015). However, the majority of human norovirus cases and outbreaks are caused by a single capsid genotype called GII.4 (Kroneman et al., 2008; de Graaf et al., 2016). The rapid evolution of viruses within the GII.4 genotype has necessitated the subdivision of this genotype into strains (Figure 3.1), each of which is named after the geographical location and year in which the strain was first detected (Siebenga et al., 2009; Parra et al., 2017). The other genogroups are associated with infection of different animal species: GIII with infection of sheep and cows, GIV.2 with infection of cats, dogs and lion, GV with infection of mice and rats and GVI and GVII with infection of dogs (Figure 1.3) (Liu et al., 1999; Ando et al., 2000; Oliver et al., 2003; Karst, 2003; Wobus et al., 2006; Park et al., 2007; Wolf et al., 2009; Martella et al., 2008, 2009; Mesquita et al., 2010; Tse et al., 2012; Vinje, 2015). While most of the GII genotypes are associated with human infection, the genotypes GII.11, GII.18 and GII.19 infect swine (Sugieda

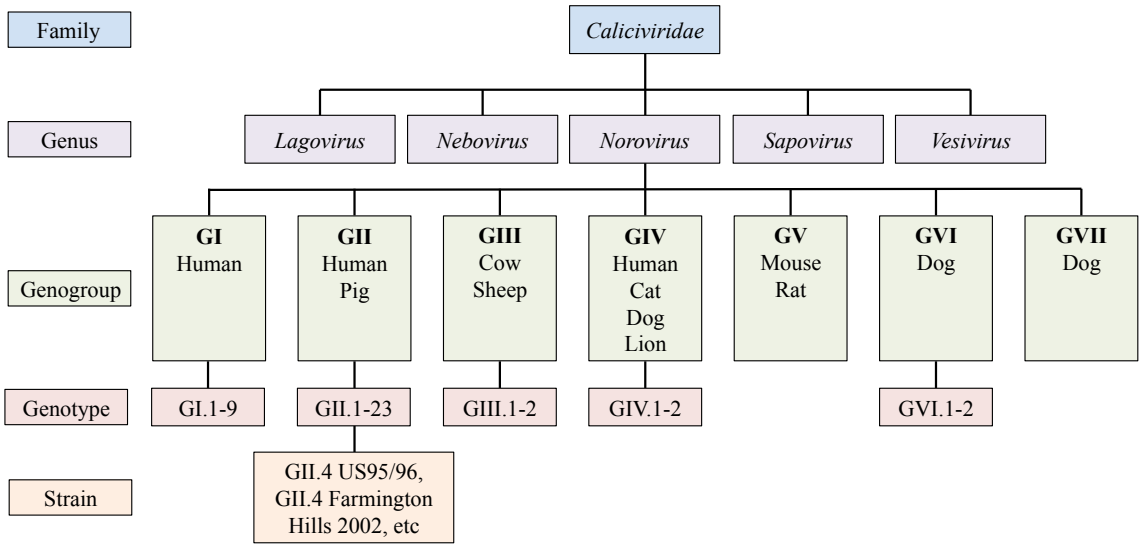


Figure 1.3: Classification of the *Caliciviridae*. The hierarchical classification of the *Caliciviridae* family is shown, with *Norovirus* being one of five accepted genera within the family. There are currently seven genogroups of norovirus, with five of these genogroups being further subdivided into multiple genotypes. The species known to be infected by each genogroup are shown. Rapid evolution within the GII.4 genotype has necessitated further subdivision of this genotype into strains.

et al., 1998; Wang et al., 2005; Wolf et al., 2009; Vinje, 2015). There have been individual studies that have reported norovirus infection of dolphin (de Graaf et al., 2017), bat (Wu et al., 2016) and sea lion (Li et al., 2011), indicating that noroviruses likely have a wide host spectrum.

1.2.3 Cell culture and tropism

Until recently, there was no reproducible and robust cell culture system for human noroviruses and MNVs were the only members of the genus that could be cultivated *in vitro* (Wobus et al., 2004; Vashist et al., 2009; Jones et al., 2014; Ettayebi et al., 2016). This has greatly hampered understanding of the viral life cycle, host-pathogen interactions and the mechanisms employed by noroviruses to infect the intestinal tract. MNVs replicate efficiently in the murine macrophage cell line RAW264.7 and in primary macrophages and dendritic cells isolated from mice, suggesting a tropism for haematopoietic cells (Wobus et al., 2004). This is supported by the detection of MNV antigen *in vivo* in cells that morphologically resemble macrophages and dendritic cells (Wobus et al., 2004; Ward et al., 2006). However, despite human norovirus antigens being detected in

cells resembling macrophages and dendritic cells in chimpanzee and immunodeficient mouse models (Bok et al., 2011; Taube et al., 2013), efforts to culture human noroviruses in these cell types have been unsuccessful (Lay et al., 2010).

However, recently B cells have been identified as a target cell for both human and murine noroviruses, leading to the development of a human norovirus cell culture system employing the BJAB B cell line (Jones et al., 2014, 2015). Interestingly, replication of a GII.4 Sydney 2012 virus within this system was decreased upon stool filtration but restored by co-culture with enteric bacteria expressing histo blood group antigens (HBGAs), suggesting that enteric bacteria may act as a stimulatory factor for norovirus infection (Jones et al., 2014, 2015). Infection of B cells *in vivo* is supported by the detection of MNV genomes and nonstructral protein in B cells of infected mice (Jones et al., 2014) and by a decrease in viral titre in mice (Jones et al., 2014) and humans (Brown et al., 2016a) lacking B cells. However, patients lacking B cells can still exhibit high viral loads, demonstrating infection of another tissue type must occur *in vivo* (Brown et al., 2016a).

A recent study detected the presence of the VP1 capsid protein, VPg and RdRp in enterocytes from the duodenum and jejunum of immunocompromised transplant patients (Karandikar et al., 2016). Importantly, the detection of nonstructural proteins suggests productive infection of enterocytes occurs *in vivo*. This is also supported by a recently developed cell culture system that has demonstrated human norovirus replication in human intestinal enteroids (HIEs) (Ettayebi et al., 2016). This cell culture system is developed from stem cells isolated from human intestinal crypts and forms a multicellular, differentiated and physiologically active culture containing enterocytes, enteroendocrine cells, goblet cells and Paneth cells (Sato et al., 2011; Ettayebi et al., 2016). Replication of human noroviruses from genotypes GI.1, GII.3, GII.4 and GII.17 has been demonstrated in this system, with enterocytes being the only cell type within the system to be infected (Ettayebi et al., 2016). Interestingly, while GII.4 viruses replicate within HIEs without the need for additional factors, GI.1, GII.3 and GII.17 viruses require a nonproteinaceous component of bile to support replication, with this component likely acting upon the cells to support infection. The addition of bile also enhanced GII.4 infection. Both GII.3 and GII.4 viruses replicate in HIEs from the duodenum, jejunum and ileum, suggesting noroviruses can productively infect enterocytes in each region of the small intestine. Stud-

ies examining GII.4 norovirus infection of gnotobiotic pigs and GIII.1 infection of calves have also found evidence of *in vivo* infection of enterocytes (Cheetham et al., 2006; Otto et al., 2011).

The VP1 capsid protein was also detected in macrophages, T cells and dendritic cells in the lamina propria of immunocompromised transplant patients, although VPg and RdRp were not detected in T cells or dendritic cells and the presence of these nonstructural proteins in macrophages was likely due to phagocytosis of a productively infected enterocyte (Karandikar et al., 2016). Together, the current evidence suggests that enterocytes are a major site of human norovirus replication *in vivo*, with additional replication in B cells and potentially in other immune system cell types (Jones et al., 2014; Ettayebi et al., 2016; Karandikar et al., 2016; Green, 2016; Brown et al., 2016a).

1.2.4 VLPs and expression systems

Due to the lack of a human norovirus cell culture system until recently (Jones et al., 2014; Ettayebi et al., 2016), many laboratory assays have taken advantage of the fact that the VP1 capsid protein self-assembles to form virus like particles (VLPs) upon *in vitro* expression (Baric et al., 2002; Donaldson et al., 2008). Two recombinant expression systems have been developed that enable the production of large quantities of VLPs that are antigenically highly similar to native particles but lack the viral genome: the baculovirus system and the Venezuelan equine encephalitis (VEE) replicon system (Jiang et al., 1992; Baric et al., 2002; Donaldson et al., 2008, 2010). In the baculovirus system, wild type baculovirus DNA is co-transfected with a transfer vector DNA containing the norovirus capsid gene as a cDNA copy (Jiang et al., 1992). In the VEE replicon system, the VEE structural genes are provided *in trans* but are replaced in the genome by the norovirus capsid gene under the control of a subgenomic 26S promoter (Baric et al., 2002). The VEE expression system has the advantage of employing mammalian rather than insect cells which may result in more physiological post-translational modifications (Baric et al., 2002). VP2 increases the stability of VLPs but is not required for VLP formation (Bertolotti-Ciarlet et al., 2003).

1.2.5 The norovirus replication cycle

The first stage of the viral replication cycle is entry into a susceptible target cell. Viral entry is typically a multistep process that involves binding of the viral particle to one or more attachment factors to concentrate the viral particles on the cell surface followed by interaction with one or more receptors that promote uptake into the cell (Mercer et al., 2010; Bartnicki et al., 2017). While norovirus entry is incompletely understood, a similar multistep entry process has been hypothesised (Figure 1.4) (Bartnicki et al., 2017). The attachment factors for human noroviruses are HBGAs, small carbohydrate molecules that are expressed on the surface of intestinal epithelial cells and are secreted into body fluids, including saliva (Figure 1.4) (Lindesmith et al., 2008; Thorne and Goodfellow, 2014). HBGAs can be found on a variety of N- and O-linked glycoproteins and on several groups of glycosphingolipids (GSLs), the latter of which are major components of cellular membranes (Taube et al., 2010). Noroviruses bind to HBGAs from the A/B/H and Lewis antigens in a genotype- and strain-specific manner (Lindesmith et al., 2008; Singh et al., 2015). The expression of HBGAs is under the control of fucosyl and glycosyltransferases encoded by the *FUT2*, *FUT3* and *ABH* genes (Figure 1.5) (Ruvoën-Clouet et al., 2013; Singh et al., 2015). Polymorphisms in these genes result in differential HBGA expression in saliva and on the surface of gastrointestinal cells. The frequency of such polymorphisms varies greatly depending on ethnicity (Lindesmith et al., 2003; Nordgren et al., 2016). Of particular importance for norovirus infection are several inactivating polymorphisms in the *FUT2* gene, including the G428A nonsense mutation which is homozygous in roughly 20% of the European and North American populations (Lindesmith et al., 2003; Le Pendu et al., 2006; Carlsson et al., 2009; Kindberg and Svensson, 2009; Kambhampati et al., 2015). The FUT2 enzyme adds a fucose moeity to carbohydrate chains and is required for the expression of multiple HBGAs on the surface of gastrointestinal cells (Figure 1.5). Individuals with a nonfunctional FUT2 enzyme only express the Le^A and Le^X HBGAs and are termed non-secretors, while individuals with a functional FUT2 enzyme are termed secretors and express a greater range of HBGAs (Figure 1.5) (Lindesmith et al., 2003; Ruvoën-Clouet et al., 2013). Non-secretor individuals account for 5%-50% of individuals in different populations worldwide (Nordgren et al., 2016). Different norovirus genotypes are associated with different host susceptibility patterns.

For example, the GI.1 Norwalk virus is unable to infect non-secretors as it binds only to HBGA types A and H and Lewis antigen B (Lindesmith et al., 2003; Le Pendu et al., 2006). Multiple studies have also demonstrated non-secretors are resistant to infection with the globally dominant GII.4 genotype (Bucardo et al., 2009; Carlsson et al., 2009; Nordgren et al., 2013; Liu et al., 2014; Currier et al., 2015; Lopman et al., 2015; Ettayebi et al., 2016). However, certain GII.4 strains can bind non-secretor HBGAs *in vitro* (Singh et al., 2015) and rare cases of GII.4 infection of non-secretors have been reported (Carlsson et al., 2009; Nordgren et al., 2013). There are, however, multiple genotypes that can infect both secretors and non-secretors, including GI.3 and GII.2 (Lindesmith et al., 2005; Rockx et al., 2005; Nordgren et al., 2010), although non-secretor-specific genotypes have not been identified. While there is strong support for HBGAs acting as norovirus attachment factors, the existence of additional attachment factors has been suggested due to the identification of strains with weak affinity for all tested HBGAs (Lindesmith et al., 2008, 2012a), although the specific molecules are yet to be described.

The HBGA-binding site has been mapped in the capsid of several norovirus genotypes (Figure 1.5) (Cao et al., 2007; Singh et al., 2015; Koromyslova et al., 2015; Singh et al., 2016b). Capsid dimerisation is essential for HBGA-binding, with the binding site being close to the dimerisation interface and formed by residues in both monomers (Cao et al., 2007). Therefore each capsid dimer contains two HBGA-binding sites (Singh et al., 2015). Five capsid residues are key for the binding of each tested HBGA with GII.4 VLPs from multiple strains: T344, R345 and D374 in one monomer and G443 and Y444 in the other monomer (Figure 1.5) (Cao et al., 2007; Singh et al., 2015). These residues interact with the fucose in both ABH and Lewis HBGAs (Singh et al., 2015). Additional residues are important for binding to specific HBGAs, including S343, Y390, C441, S442 and the loop containing residues 391-395 (Figure 1.5) (Donaldson et al., 2008; Singh et al., 2015). Specifically, the 391-395 loop is repositioned to enable binding to HBGA type B, Le^B and Le^Y (Singh et al., 2015). Additionally, residue T338 is predicted to be essential for HBGA binding through hydrogen bonds formed between this residue and R345 (Cao et al., 2007; Donaldson et al., 2008). The mode of interaction between the capsid and HBGA receptor differs between GI and GII noroviruses; in the GI.1 Norwalk virus capsid, residues D327, H329, S377 and S380 interact with the N-acetylglucosamine of the HBGA type A (Bu et al., 2008).

Capsid site	Function	Additional information	Reference
294	Blockade epitope A		Lindesmith et al. (2012b)
296	Blockade epitope A/site A		Allen et al. (2008); Lindesmith et al. (2012b)
297	Blockade epitope A/site A		Allen et al. (2008); Lindesmith et al. (2012b)
298	Blockade epitope A/site A		Allen et al. (2008); Lindesmith et al. (2012b)
310	Access to 'universal epitope'		Lindesmith et al. (2014)
316	Access to 'universal epitope'	Glutamic acid within GII.4	Lindesmith et al. (2014)
333	Epitope B		Lindesmith et al. (2012b)
338	HBGA-binding	Hydrogen bonds to residue R345, threonine within GII.4	Cao et al. (2007)
340	Epitope C		Lindesmith et al. (2012b)
343	HBGA-binding site, specific	Serine within GII.4	Cao et al. (2007); Singh et al. (2015)
344	HBGA-binding site, required	Threonine within GII.4	Cao et al. (2007); Singh et al. (2015)
345	HBGA-binding site, required	Arginine within GII.4	Cao et al. (2007); Singh et al. (2015)
355	May contribute to blockade epitope E		Lindesmith et al. (2012b)
356	May contribute to blockade epitope E		Lindesmith et al. (2012b)
357	May contribute to blockade epitope E		Lindesmith et al. (2012b)
368	Blockade epitope A		Lindesmith et al. (2012b)
372	Blockade epitope A		Lindesmith et al. (2012b)
373	Blockade epitope A		Debbink et al. (2013); Allen et al. (2014)
374	HBGA-binding site, required	Aspartic acid within GII.4	Cao et al. (2007); Singh et al. (2015)
376	Epitope C		Lindesmith et al. (2012b)
382	Epitope B		Lindesmith et al. (2012b)
390	HBGA-binding site, specific	Tyrosine within GII.4	Singh et al. (2015)
391-395 loop	HBGA binding site, specific	Repositioning of this loop is required to bind HBGA type B, Le^B and Le^Y	Singh et al. (2015)
393	Blockade epitope D/site B		Allen et al. (2008); Lindesmith et al. (2012b)
394	Blockade epitope D/site B		Allen et al. (2008); Lindesmith et al. (2012b)
395	Blockade epitope D/site B		Allen et al. (2008); Lindesmith et al. (2012b)
407	Blockade epitope E		Lindesmith et al. (2012b)
412	Blockade epitope E		Lindesmith et al. (2012b)
413	Blockade epitope E		Lindesmith et al. (2012b)
441	HBGA-binding site, specific	Cysteine within GII.4	Cao et al. (2007); Singh et al. (2015)
442	HBGA-binding site, specific	Serine within GII.4	Cao et al. (2007); Singh et al. (2015)
443	HBGA-binding site, required	Glycine within GII.4	Cao et al. (2007); Singh et al. (2015)
444	HBGA-binding site, required	Tyrosine within GII.4	Cao et al. (2007); Singh et al. (2015)
484	Access to 'universal epitope'	Arginine within GII.4	Lindesmith et al. (2014)
493	Access to 'universal epitope'	Lysine within GII.4	Lindesmith et al. (2014)

Table 1.1: Summary of epitopes and HBGA-binding sites in the GII.4 capsid. Each site within the GII.4 capsid suggested to form part of an epitope region or contribute to HBGA-binding is shown. Sites listed as 'HBGA-binding site, required' are required for binding to all HBGAs, while sites listed as 'HBGA-binding site, specific' are only required for binding to specific HBGAs. Site 338 is essential for HBGA-binding due to a hydrogen bond formed by this residue to R345. Numbering relative to the post-US95/96 GII.4 capsid (i.e. after the insertion at site 394).

The interaction between the capsid P2 domain and HBGAs is thought to result in clustering of the GSLs to form a lipid raft (an organised lipid microdomain), leading to local invagination of the membrane (Bartnicki et al., 2017). Binding of norovirus to HBGAs is not sufficient to mediate entry into target cells and, while no receptor for human norovirus has yet been described, two groups recently independently identified CD300lf as a proteinaceous receptor for MNV (Orchard et al., 2016; Haga et al., 2016). CD300lf knockout mice are resistant to MNV infection and the expression of murine CD300lf in cells from other species makes these usually non-permissive cell types permissive to MNV infection (Orchard et al., 2016; Haga et al., 2016). While CD300ld was also demonstrated to be a receptor for MNV (Orchard et al., 2016; Haga et al., 2016), the lack of viral replication in CD300lf knockout mice suggests that CD300lf is the primary MNV receptor (Figure 1.4) (Orchard et al., 2016). The entry process of norovirus into cells is dependent on dynamin, cholesterol and ceramide (Thorne and Goodfellow, 2014; Bartnicki

et al., 2017). A recently proposed model of norovirus entry suggests that binding to one or more GSLs enables attachment of the viral particle to the cell surface. The GSLs are either already in lipid rafts or are recruited into a new lipid raft where the CD300 protein receptors are located. The lipid rafts are dependent on cholesterol and interactions within the lipid raft may result in invagination of the plasma membrane followed by dynamin II-mediated scission of the invagination (Figure 1.4) (Bartnicki et al., 2017).

After entry into host cells, the viral capsid disassembles in a process known as uncoating, resulting in release of the VPg-linked viral genome into the cytoplasm where it acts as the template for the 'pioneer round' of protein translation (Figure 1.4) (Thorne and Goodfellow, 2014). Noroviruses do not encode their own translational machinery and translation is therefore dependent on the recruitment of host cell factors and ribosomes. VPg acts as a substitute for 5' cap that is found on cellular mRNAs and acts in the same way to recruit cellular eukaryotic translation initiation factors (eIFs), including eIF3, eIF4e, eIF4G and cap binding protein (Figure 1.4) (Thorne and Goodfellow, 2014; Chung et al., 2014; Alhatlani et al., 2015). Therefore linkage of the viral genome to VPg is essential for infectivity. Binding of these cellular factors by VPg results in recruitment of the 43S ribosomal pre-initiation complex (Thorne and Goodfellow, 2014). Additionally, highly conserved RNA structures at the 5' and 3' ends of the viral genome interact with cellular proteins and, while the exact function(s) of most of these interactions are yet to be elucidated, they likely enable circularisation of the genome and efficient translation (Thorne and Goodfellow, 2014; Alhatlani et al., 2015). These interactions may also result in a gradual switch from translation in the early phase of infection to replication in the late phase of infection via recruitment of the cellular protein PTB; mutation of the PTB-binding site in the MNV genome resulted in viral attenuation *in vivo*, demonstrating the essential role of host cell proteins in infection (Alhatlani et al., 2015).

Translation of the VP1 capsid protein and VP2 occurs primarily from the subgenomic RNA, which is likely a strategy to enable efficient production of the high levels of these proteins required to form new viral particles (Thorne and Goodfellow, 2014). The subgenomic RNA is polycistronic and translation of VP2 occurs by a termination-reinitiation mechanism. In this mechanism, ribosomes remain associated with the RNA after termination of ORF2 translation and reinitiate translation at the start of ORF3 (Napthine et al., 2009). Cellular RNA is typically present at far higher levels than the incoming viral RNA

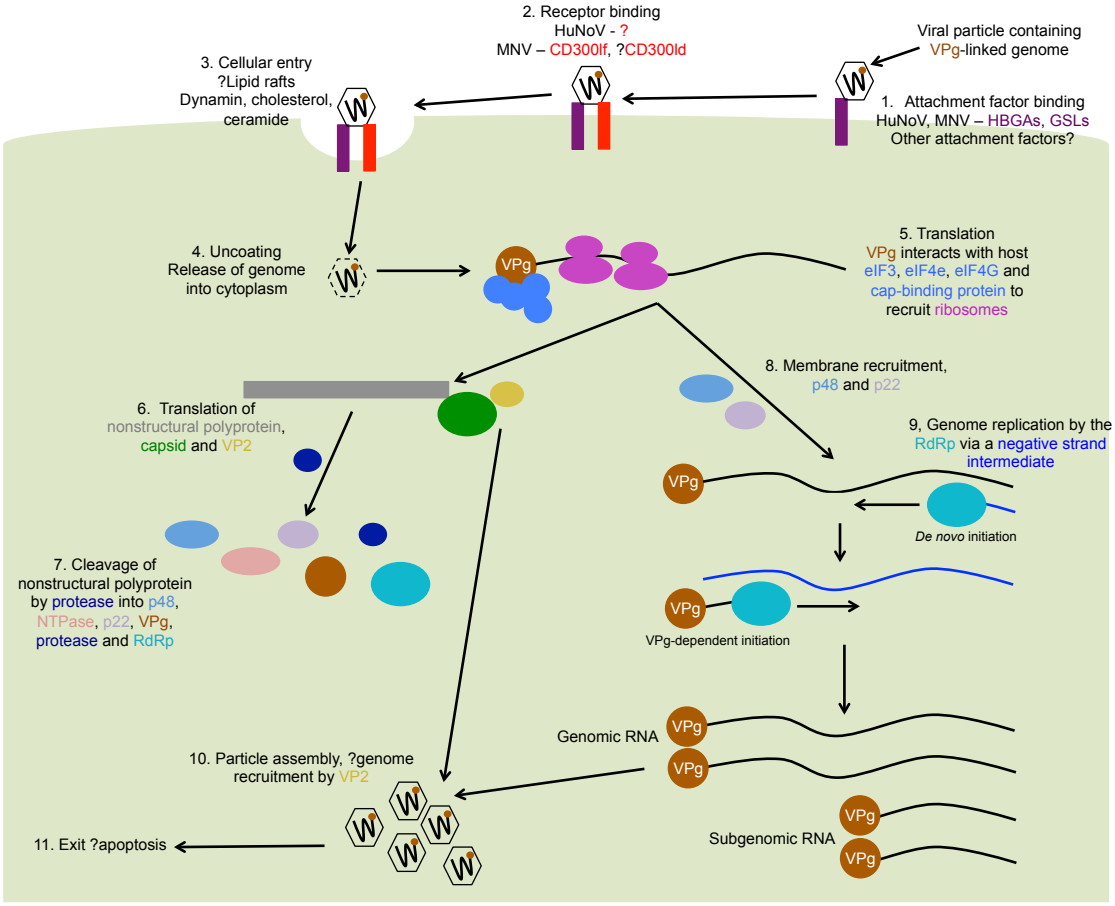


Figure 1.4: The norovirus replication cycle. Each stage in the replication cycle is described in more detail in the text. Briefly, the viral particle interacts with cellular attachment factors that enable interaction with additional receptors and entry into the cell. After cell entry, the genome is released from the viral particle and acts as the template for the 'pioneer' round of translation. This translation is carried out by host ribosomes recruited through interactions between VPg and cellular eukaryotic translation initiation factors (eIFs). The virally encoded protease cleaves the nonstructural polyprotein into six proteins. The replication of the viral genome is carried out by the virally-endoded RdRp and occurs on membranes that are recruited by p48 and p22. Genome replication occurs via a negative strand intermediate and results in the production of genomic RNA and subgenomic RNA. These RNA molecules are used for additional translation, with high levels of the capsid protein and VP2 being translated from subgenomic RNA. The capsid and VP2 assemble to form viral particles, with VP2 likely recruiting the viral genome and these viral particles are released from the cell.

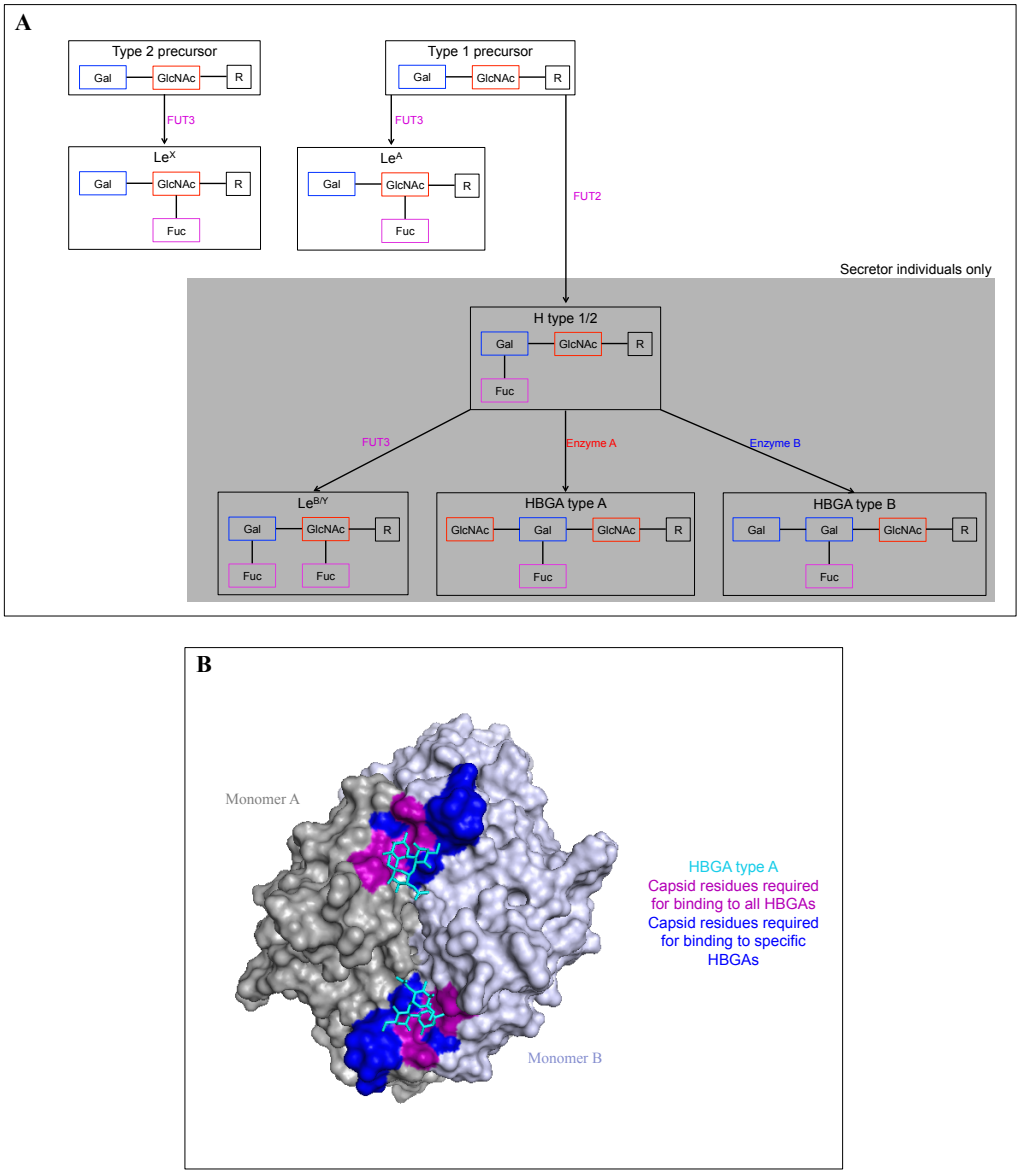


Figure 1.5: The HBGA receptors used by noroviruses as attachment factors. (**A**) The biosynthesis pathway for HBGA receptors. Gal - galactose, GlcNAc - N-acetylglucosamine, Fuc - fucose. The synthesis of HBGAs is under the control of fucosyl- and glycosyltransferases encoded by the *FUT2*, *FUT3* and *ABH* genes. The type 1 and type 2 precursor molecules differ by the nature of the link between the Gal and GlcNAc moities, with the type 1 precursor exhibiting a β1,3 link and the type 2 precursor exhibiting a β1,4 link. The enzyme that catalyses each step in the synthesis pathway is shown by the corresponding arrow and the sugar group added by each enzyme is shown in the same colour as the enzyme. Le^A and Le^X differ by only the linkages between the sugar groups, as do Le^B and Le^Y. The structure of Le^Y is not shown for simplicity. The HBGAs in the grey shaded area require the action of a functional FUT2 enzyme to be produced and so are only expressed in secretor individuals. Non-secretor individuals can only express Le^A and Le^X. The HBGAs shown are terminal structures, the rest of the chain is shown by the R group. (**B**) The HBGA-binding site is highlighted on the GII.4 Sydney 2012 dimeric P domain structure 4WZT. This is a top view of the capsid with one monomer shown in grey and the other monomer shown in blue-white. The HBGA (in this case HBGA type A) is shown in cyan. Residues T344, R345, D374, G443 and Y444 that are required for binding to all HBGAs are shown in purple. Residues S343, Y390, C441, S442 and the 391-395 loop that are required for binding to specific HBGAs are shown in blue.

and many viruses have evolved mechanisms to promote translation of viral proteins over that of cellular proteins (Firth and Brierley, 2012). A recent study demonstrated that the norovirus protease cleaves the cellular protein PABP to enable the preferential translation of viral mRNA and a reduction in translation of cellular proteins (Emmott et al., 2017b). This has the additional benefit of inhibiting translation of interferon-stimulated genes and may therefore act as a mechanism of immune evasion (Emmott et al., 2017b). The viral protease cleaves the nonstructural polyprotein in a co- and post-translational fashion to release free forms of each of the nonstructural proteins as well as stable intermediate forms (Figure 1.4) (Belliot et al., 2003; Thorne and Goodfellow, 2014; Emmott et al., 2017a). These intermediate forms can be functional, with, for example, the protease-RdRp intermediate exhibiting polymerase activity (Belliot et al., 2005). The protease active site consists of a catalytic triad of H30, E54 and C139, with E54 likely determining substrate specificity (Someya et al., 2002, 2008; Zeitler et al., 2006; Thorne and Goodfellow, 2014).

The replication of the norovirus genome occurs in association with cellular membranes, in particular those of the endoplasmic reticulum (ER), Golgi apparatus and endosomes that form the secretory pathway (Hyde et al., 2009; Thorne and Goodfellow, 2014). While the mechanisms by which cellular membranes are recruited are not completely understood, p48 (NS1/2) and p22 (NS4) have been implicated in this process, with both proteins localising to components of the secretory pathway (Figure 1.4) (Bailey et al., 2010; Hyde and Mackenzie, 2010; Thorne and Goodfellow, 2014). The ER-export signal in p22 is thought to promote the uptake of p22 into COPII-coated vesicles in transit from the ER to the Golgi, resulting in mislocalisation of COPII-coated vesicles and Golgi disassembly. As MNV NS4 does not contain this ER-export signal, it likely acts via a different mechanism. The p48 protein is thought to act via interaction with SNARE regulator vesicle associated membrane protein-associated protein A (VAP-A) and VAP-B, proteins that regulate cellular vesical transport (McCune et al., 2017). Interaction with these proteins results in Golgi disassembly and is essential for efficient replication of murine norovirus (McCune et al., 2017).

As with other positive strand RNA viruses, replication of the genome occurs via a negative-sense intermediate (Figure 1.4). There are two mechanisms by which the norovirus RdRp can initiate synthesis of a new RNA strand: *de novo* and VPg-dependent (Rohayem et al., 2006; Subba-Reddy et al., 2011). Both mechanisms have been demon-

strated *in vitro* (Rohayem et al., 2006). It has been proposed that the norovirus RdRp synthesises the negative-sense genomic and sub-genomic RNA by *de novo* initiation (Figure 1.4) (Subba-Reddy et al., 2012). The active site of the RdRp consists of residues R182, D242, Y243, D247, S300, N309, D343 and D344 (Ng et al., 2004; Zamyatkin et al., 2008) and the RdRp is thought to function as a homodimer (Högbom et al., 2009). Specific loops within the shell domain of the VP1 capsid protein enhance *de novo* initiation in a species-specific manner (Subba-Reddy et al., 2012). This has been proposed to enhance genome replication when levels of the VP1 capsid protein are low early in the replication cycle. As more copies of VP1 are present, they form higher order capsid structures which prevents RdRp-binding, thereby decreasing genome replication (Subba-Reddy et al., 2012; Thorne and Goodfellow, 2014). Species-specific interaction between p48 and RdRp also enhances RdRp activity, while VP2 interaction with RdRp inhibits RdRp activity (Subba-Reddy et al., 2011).

Synthesis of the negative-sense RNA strand results in formation of a double-stranded replicative form. Synthesis of the positive-sense genomic RNA is thought to occur via VPg-dependent initiation (Figure 1.4) (Chaudhry et al., 2006; Thorne and Goodfellow, 2014). The first step of this synthesis involves VPg nucleotidylation (also called VPg guanylation), where the RdRp covalently attaches a conserved tyrosine residue in VPg (Y27 in human norovirus and Y26 in murine norovirus) to the initiating nucleotide, which is guanine throughout the noroviruses (Subba-Reddy et al., 2011). The nucleotidylation process has been suggested to be enhanced by an element at the 3' end of the genome (Belliot et al., 2005). Two potential, and non-mutually exclusive, mechanisms by which the subgenomic RNA molecules may be transcribed have been suggested. Both positive-sense and negative-sense genomic and subgenomic RNAs have been identified in infected cells (Green et al., 2002). In the first mechanism, the RdRp reaches a termination signal at the 5' end of VP1 while transcribing the negative-sense RNA strand, thereby resulting in negative-sense subgenomic RNA molecules from which positive-sense subgenomic RNAs can be transcribed. In the second mechanism, a promoter sequence exists in the genomic negative-sense RNA at the start of VP1 and the RdRp transcribes positive-sense subgenomic RNA starting at this promoter (Thorne and Goodfellow, 2014). The existence of a conserved RNA stem loop structure close to the start of the subgenomic RNA supports this second mechanism, although whether this loop has a direct role in synthesis of the

subgenomic RNA is unknown (Simmonds et al., 2008).

Additional rounds of RNA replication and viral protein translation occur (Figure 1.4). That the VP1 capsid protein self-assembles into virus-like particles *in vitro* suggests that cellular factors are unlikely to be required for formation of the viral particle (Baric et al., 2002). VP2 is located on the interior face of the viral particle and contains a basic domain that may recruit the viral genome into the particle (Vongpunsawad et al., 2013). Such a basic domain is not present in the VP1 capsid protein. Alternatively, it is possible that an interaction between VPg and either VP1 or VP2 may recruit the genome into the particle and a possible interaction between VPg and VP1 has been observed in the related feline calicivirus (Kaiser et al., 2006). While studies of the mechanism(s) employed by noroviruses to exit host cells are largely lacking, apoptosis is induced by murine norovirus and apoptotic cells have been noted in intestinal biopsies from infected patients (Bok et al., 2009; Furman et al., 2009; Karandikar et al., 2016). Damaged sites in intestinal biopsies correlate with the expression of viral antigens (Karandikar et al., 2016). Induction of apoptosis by MNV is associated with downregulation of the cellular pro-survival factor survivin (Bok et al., 2009; Furman et al., 2009) and inhibition of apoptosis reduces the production of MNV particles, suggesting the importance of this mechanism in the viral replication cycle (Figure 1.4) (Furman et al., 2009).

1.3 Transmission and epidemiology

1.3.1 Norovirus transmission

Noroviruses transmit exceptionally efficiently between susceptible individuals via the faecal-oral or oral-oral route (de Graaf et al., 2016). Part of the reason explaining this success may be a low infectious dose, with the estimated ID50 (number of viral particles required to infect 50% of exposed individuals) being as low as 18 viral particles for the GI.1 Norwalk virus (Teunis et al., 2008). Additionally, the same study suggested that exposure to a single virus particle results in a 50% probability of infection (Teunis et al., 2008). In context, a single gram of faeces from an infected patient can contain up

to 10^{11} virus particles, with the estimated median peak of shedding containing $9.5x10^{10}$ genome copies per gram of faeces (Atmar et al., 2008). Therefore just one gram of faeces can contain more than 5 billion infectious doses (Hall, 2012). However, more recent estimates have suggested the Norwalk virus ID50 to be 1320 genome equivalents (95% CI 440-3760) (Atmar et al., 2014). To date, no study has estimated the ID50 of the dominant GII.4 genotype. Given the high prevalence of GII.4, it is possible that the infectious dose for this genotype may be lower than that of other genotypes. Noroviruses are thought to withstand freezing, heating and many commonly used detergents, although the lack of a cell culture system for human noroviruses has made it difficult to assess whether treated viruses remain infectious (Donaldson et al., 2008; Hall, 2012; Lopman et al., 2012). MNV and feline calicivirus have therefore been used as surrogates for human norovirus and have been demonstrated to be inactivated by UV, high temperatures and high pressure (Lopman et al., 2012). The recent development of human norovirus culture systems (Jones et al., 2014; Ettayebi et al., 2016) may enable determination of whether the same is true for human noroviruses. It has been demonstrated that norovirus can persist on surfaces for up to two weeks and even longer persistence has been demonstrated in water, where virus particles spiked into ground water remained infectious in humans for at least 61 days (Seitz et al., 2011; Lopman et al., 2012). Intact viral particles containing genomic RNA remained detectable in ground water for more than three years, although whether these particles were infectious was not determined (Seitz et al., 2011). This long environmental persistence is exacerbated by the ability of vomiting to aerosolise the virus, thereby spreading the virus to a large area (Lopman et al., 2012). While the highest levels of contamination occur on surfaces that have been in direct contact with faeces or vomit, contamination of a large variety of objects has been demonstrated and vomiting is likely to accelerate the spread of the virus (Cheesbrough et al., 2000; Gallimore et al., 2006, 2008). In addition, the short incubation time coupled with the long period of shedding after the resolution of symptoms means that infected patients have a long period of time during which they are potentially infectious (Patel et al., 2009; Lee et al., 2013; Lopman et al., 2012; Teunis et al., 2015). However, symptomatic patients have been suggested to be responsible for most transmission within norovirus outbreaks (de Graaf et al., 2016).

While most transmission in outbreaks and sporadic disease is thought to occur through direct person to person transmission (Lopman et al., 2012), noroviruses can be

transmitted indirectly via faecal or vomit contamination of food, water, the environment or fomites. However, outbreaks can involve multiple transmission routes and can be extended and perpetuated by contaminated fomites (Lopman et al., 2012). Outbreaks also frequently occur through contaminated food and these outbreaks can be large and international due to the global nature of food production (Lopman et al., 2012; de Graaf et al., 2016). Different genotypes appear to exhibit different preferences for different transmission routes, with a greater proportion of non-GII.4 genotype outbreaks being foodborne compared with GII.4 genotype outbreaks (Verhoef et al., 2015). In total, roughly 14% of norovirus outbreaks are attributed to food (Verhoef et al., 2015). The contamination of food products can occur through irrigation with contaminated sewage or exposure of shellfish to faecal discharge, but contamination by food handlers is likely the main mechanism (Lopman et al., 2012; de Graaf et al., 2016). Foodborne outbreaks are commonly contaminated with multiple genotypes, providing a mechanism by which recombination may occur between different strains (de Graaf et al., 2016). While numerous waterborne outbreaks have been reported, these likely account for only a small percentage of total norovirus outbreaks (Lopman et al., 2012).

1.3.2 Seasonality

Norovirus cases and outbreaks exhibit peaks in the winter months and troughs in summer months in temperate regions of the Northern hemisphere (Ahmed et al., 2013). While seasonality in Oceania has been less clear, a recent study suggested norovirus outbreaks peak during the southern hemisphere winter (Eden et al., 2014), similar to temperate countries in South America (da Silva Poló et al., 2016). Of the limited studies carried out in Africa, no clear seasonal pattern of infection has been identified (Mans et al., 2016).

1.3.3 Norovirus epidemiology

While there are close to 30 genotypes known to infect humans, the vast majority of both human cases and human outbreaks of norovirus are caused by a single capsid genotype, termed GII.4, with this genotype causing up to 82% of outbreaks annually

(Siebenga et al., 2007; Kroneman et al., 2008; Siebenga et al., 2009). While the GII.3 genotype is prevalent in children (Boon et al., 2011; Brown et al., 2016c), the other non-GII.4 genotypes are typically rare (Parra et al., 2017). The GII.4 genotype has also caused six pandemics of gastroenteritis since the mid-1990s (Figure 1.6) (Siebenga et al., 2009; Bull and White, 2011; van Beek et al., 2013). Each of the pandemics has been associated with a phylogenetically distinct GII.4 strain and pandemics have occurred in the Northern hemisphere winters of 1995-1996 (US95/96 strain), 2002-2003 (Farmington Hills 2002), 2004-2005 (Hunter 2004), 2006-2007 (Den Haag 2006), 2009-2010 (New Orleans 2009) and 2012-2013 (Sydney 2012) (Figure 1.6) (Noel et al., 1999; Lopman et al., 2004a; Bull et al., 2006; Tu et al., 2008; Vega et al., 2011; van Beek et al., 2013; Siebenga et al., 2007, 2009). Additional epidemic GII.4 strains have been identified that either cause outbreaks within a more restricted geographical region, or that cause widespread but only sporadic outbreaks, including the Lanzhou 2002, Kaiso 2003, Asia 2003, Yerseke 2006, Osaka 2007 and Apeldoorn 2007 strains (Figure 1.6) (Siebenga et al., 2009; Eden et al., 2014). The first GII.4 pandemic in 1995-1996 coincided with an increase in norovirus outbreaks in USA, Europe and Australia and the Farmington Hills 2002 pandemic was associated with an unprecedented number of norovirus outbreaks (Wright et al., 1998; Lopman et al., 2004a; Bull and White, 2011). While the first pandemic coincided with an increase in norovirus surveillance, raising the possibility of additional earlier pandemics that remained undetected, phylodynamic analysis suggests that there was an increase in GII.4 norovirus prevalence in the mid-1990s (Siebenga et al., 2010). The most recent common ancestor of the GII.4 capsid has been dated to 1966, suggesting that this genotype was present for at least 30 years prior to the first pandemic (Bok et al., 2009). There are very few studies of norovirus molecular epidemiology prior to the emergence of the first GII.4 pandemic. However, each of the studies carried out to our knowledge to date that have genotyped norovirus samples collected in the 1970s and 1980s have suggested that GII.4 was one of a number of lower prevalence genotypes at this time (Boon et al., 2011; Siqueira et al., 2017; Mori et al., 2017b,a). GII.4 was only present in 17% of norovirus-positive samples collected from children in a Washington hospital between 1974 and 1991 (Bok et al., 2009; Boon et al., 2011), while in Northern Brazil GII.4 was detected at comparable frequencies to GII.6 and GII.7 between 1982 and 1986, to GII.3, GII.6 and GII.7 between 1990 and 1992 and to GII.3 between 1992 and 1994 (Siqueira

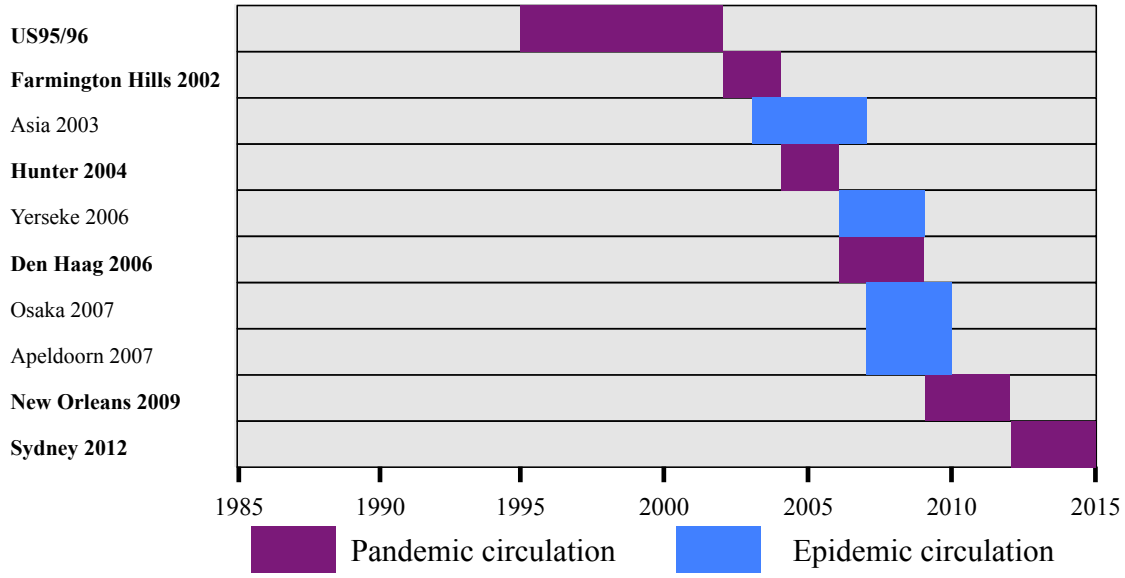


Figure 1.6: Pandemic and epidemic GII.4 strains. The main period of pandemic (purple) or epidemic (blue) circulation is shown for each of the major GII.4 strains. The pandemic strain labels are in bold font.

et al., 2017). Importantly, following the emergence of US95/96, this strain dominated norovirus-positive samples in Northern Brazil between 1998 and 2000 (Siqueira et al., 2017). GII.4 was identified in only 6.9% of norovirus foodborne outbreaks in Tokyo between 1966 and 1983 (Mori et al., 2017b) and in 2.3% of norovirus sporadic gastroenteritis cases in Tokyo between 1985 and 1992 (Mori et al., 2017a). This lower historical prevalence of GII.4 is also supported by a study examining serological samples collected from children in the Netherlands in 1963, 1983 and 2006/2007 (Van Beek et al., 2016). Here, the prevalence of sera reactive against GI viruses decreased through time, while sera reactive against GII.4 increased in prevalence in the 2006/2007 serum samples compared with those from 1963 and 1983 (Van Beek et al., 2016). However, this study has the caveat that serum reactivity against GII.4 was measured against a virus collected in 1995 and if the virus had undergone antigenic drift between 1963 and 1995, the ability of the 1963 serum to bind to the 1995 virus would have been reduced. However, the current evidence suggests that GII.4 first became dominant in the mid-1990s, correlating with an increase in norovirus activity (Siebenga et al., 2010).

The mechanisms underlying the high prevalence of GII.4 in recent times are not well understood, but previous proposals include broader HBGA binding (Lindesmith et al., 2008), a faster mutation rate enabling more efficient transmission and/or rapid evasion

of host immunity (Bull et al., 2010; Arias et al., 2016) and reduced capsid constraints resulting in a greater ability to accomodate change (Boon et al., 2011; Donaldson et al., 2010; Parra et al., 2017). As discussed previously, GII.4 viruses are not thought to infect nonsecretors individuals, with rare exceptions (Carlsson et al., 2009). The US95/96 pandemic strain was demonstrated to bind to secretor-positive sera from individuals of blood types A, B and O, representative of roughly 80% of the European and North American populations (Lindesmith et al., 2008; Singh et al., 2015). The US95/96 capsid also binds to the A, B, H3, Le^B and Le^Y HBGAs (Lindesmith et al., 2008). This is a greater range of binding than in the majority of other genotypes, suggesting that GII.4 may be highly prevalent due to the ability to infect a greater proportion of individuals (Lindesmith et al., 2008; Donaldson et al., 2010). However, the Hunter 2004 pandemic strain does not bind efficiently to any tested HBGA under experimental conditions (Lindesmith et al., 2012a) and therefore broad HBGA-binding does not appear to be universal in pandemic GII.4 strains (Lindesmith et al., 2008). Additionally, broad HBGA-binding is not sufficient for high prevalence, as the rarely detected GII.10 capsid exhibits broad HBGA binding (Hansman et al., 2011).

The GII.P4 RdRp was demonstrated to have a higher mutation rate in vitro than the GII.Pb (now reclassified as GII.P21 (Kroneman et al., 2013)) and GII.P7 RdRps (Bull et al., 2010). Here, the number of transversion and transition mutations introduced upon transcription from a poly-C template RNA was several fold higher for GII.P4. It was suggested that this higher mutation rate enables more efficient generation of viral diversity (Bull et al., 2010). It was recently demonstrated that RdRp fidelity is an important component of MNV fitness, with a mutation conferring increased RdRp fidelity reducing the efficiency of transmission (Arias et al., 2016). While there has been no study to date examining the influence of RdRp fidelity on human norovirus transmission, these studies together suggest that the higher mutation rate of the GII.P4 RdRp may enable more efficient transmission.

In a study comparing GII.3 and GII.4 capsid evolution over a similar time period, both genotypes exhibited an accumulation of nucleotide change through time (Boon et al., 2011). However, while GII.4 exhibited an accumulation of amino acid change through time, the GII.3 capsid remaining relatively constant at the amino acid level (Boon et al., 2011). A recent study examined a greater number of norovirus genotypes and found

that the non-GII.4 genotypes remain 'static' through time and segregate into only a small number of distinct variants, each of which circulates with few amino acid changes for long time periods (Parra et al., 2017). For example, the GII.6 genotype segregates into three variants that have each circulated with few nonsynonymous changes for more than 40 years (Parra et al., 2017). In contrast, GII.4 exhibits rapid change at the amino acid level with a greater number of variants that each persists for a much shorter period of time (Siebenga et al., 2007; Boon et al., 2011; Parra et al., 2017). This has led to suggestions that GII.4 is genetically robust and can accumulate amino acid changes while maintaining important functions, while the non-GII.4 genotypes are genetically fragile and cannot tolerate amino acid change (Parra et al., 2017). This hypothesis therefore suggests that the GII.4 capsid can structurally tolerate changes in a way that the capsid from other genotypes cannot. A similar mechanism has been proposed for the higher prevalence of GII viruses compared with GI viruses (Donaldson et al., 2010), where the GII capsid is larger with more flexible surface loops and can therefore tolerate a greater amount of change than the GI capsid. However, as the GII.4 capsid is not larger than the capsid of other GII genotypes, size alone cannot explain the higher prevalence of GII.4.

1.3.4 Emergence of GII.4 pandemic strains

Since the first GII.4 pandemic in 1995, a new pandemic GII.4 strain has emerged within the human population every 2-7 years (Figure 1.6) (Siebenga et al., 2009; van Beek et al., 2013). The emergence of a new pandemic strain is often, but not always, associated with an increase in norovirus outbreaks (Lopman et al., 2004a; Siebenga et al., 2009; Allen et al., 2014). The onset of a new pandemic is marked by rapid global strain replacement, with the new pandemic strain dominating outbreaks worldwide within several months of its first detection (Eden et al., 2014). For example, the most recent pandemic strain Sydney 2012 was first detected in Australia in March 2012 and between November 2012 and January 2013 replaced the previous pandemic strain New Orleans 2009 as the dominant cause of norovirus outbreaks in Europe, North America, Asia and Oceania (Allen et al., 2014; Giammanco et al., 2013; Hasing et al., 2013; Vega et al., 2014a; Kim et al., 2013; White, 2014). The pattern of GII.4 strain emergence is consistent with

epochal evolution, with periodic rapid emergence of a new strain followed by a period of stasis where that strain dominates (Siebenga et al., 2007; Lindesmith et al., 2008).

The emergence of a new pandemic GII.4 strain is thought to be driven by escape from herd immunity raised against the preceding pandemic strain (Lindesmith et al., 2008; Cannon et al., 2009; Lindesmith et al., 2012a; Debbink et al., 2012a; Lindesmith et al., 2013; Debbink et al., 2013). Multiple studies have demonstrated the occurrence of amino acid substitutions through time within the GII.4 genotype, with the majority of these substitutions localising to the P2 domain (Siebenga et al., 2007; Lindesmith et al., 2008; Siebenga et al., 2009; Bull et al., 2010; Lindesmith et al., 2011; Parra et al., 2017). Genetic changes within the capsid have been demonstrated to result in antigenic variation through time (Lindesmith et al., 2011), with newly emerging GII.4 strains exhibiting antigenic differences relative to their temporal predecessor (Lindesmith et al., 2013; Debbink et al., 2013). These antigenic differences were demonstrated using a combination of mouse monoclonal antibodies (mAbs) (Lindesmith et al., 2008, 2011), human mAbs (Lindesmith et al., 2012a), mouse polyclonal sera (Lindesmith et al., 2008) and human polyclonal sera (Lindesmith et al., 2008, 2013; Debbink et al., 2013). Antibody binding to a VLP is not indicative of neutralisation of infection. In the absence of a human norovirus cell culture system, a surrogate neutralisation assay has been developed that measures the ability of mAbs or sera to block the binding of a VLP to a HBGA binding partner (Lindesmith et al., 2008). Antibodies that block such an interaction are termed blockade antibodies and the presence of these antibodies in sera correlates with protection from infection in chimpanzees (Bok et al., 2011) and humans (Reeck et al., 2010). This surrogate neutralisation assay has been demonstrated to be highly sensitive and can distinguish VLPs too similar to be separated by antibody binding alone (Lindesmith et al., 2008).

Epitopes within the GII.4 genotype were initially proposed by identification of rapidly evolving regions on the surface of the viral particle (Figure 1.7) (Allen et al., 2008; Lindesmith et al., 2012a). Three blockade epitopes (so called because changes at these sites alter binding of blockade antibodies) termed A, D and E have been identified (Figure 1.7) and experimentally verified by switching of these epitopes between GII.4 strains (Lindesmith et al., 2012a; Debbink et al., 2012a; Lindesmith et al., 2013; Debbink et al., 2013). Epitope A is a conformational epitope consisting of residues 294, 296-298, 368 and 372 and has been verified as a blockade epitope in multiple GII.4 strains (Linde-

smith et al., 2012a; Debbink et al., 2012a; Lindesmith et al., 2013; Debbink et al., 2013). This epitope has been proposed to be immunodominant and is estimated to account for between 40 and 55% of the total blockade antibody response (Lindesmith et al., 2013; Debbink et al., 2013). Previous studies have suggested that epitope A actually consists of several overlapping epitopes and have demonstrated that different antibodies can interact with different residues within this region (Lindesmith et al., 2013). Epitope D consists of residues 393-395 and therefore coincides with the 391-395 loop important for binding to certain HBGAs (Lindesmith et al., 2012a; Singh et al., 2015). Correspondingly, changes within epitope D have been demonstrated to have the ability to alter antigenicity and HBGA-binding (Lindesmith et al., 2008, 2012a). In particular, substitutions at site 393 can alter binding to HBGA type B, while substitutions at site 395 can alter binding to HBGA type A (Lindesmith et al., 2008). Interestingly, changes within epitope D can therefore increase the size of the susceptible population by two mechanisms; evasion of existing immunity or infection of a new part of the population (Lindesmith et al., 2008; Donaldson et al., 2008). Epitope E is a conformational epitope consisting of residues 407, 412 and 413 and was initially identified as a Farmington Hills 2002 blockade epitope (Lindesmith et al., 2012b). Residues 355-357 also likely contribute to this epitope (Lindesmith et al., 2012b). Two additional epitopes, B and C, have been proposed based on their high variability and surface location, but are yet to be confirmed experimentally (Lindesmith et al., 2012a). Epitopes B and C consist of residues 333 and 382 and residues 340 and 376, respectively (Lindesmith et al., 2012a). Each of these five epitopes likely also includes additional sites close to the variable sites (Lindesmith et al., 2012a). Importantly, there is at least one more blockade epitope present in the GII.4 capsid that has not been identified to date (Lindesmith et al., 2012a). Additionally, previous studies have identified blockade epitopes on the basis of HBGA binding, raising the possibility of additional blockade epitopes that influence binding to the as yet undiscovered major norovirus receptor.

One of the human mAbs (NVB 71.4) isolated by Lindesmith et al. (2012a) blocks the interaction of a wide range of GII.4 VLPs collected from 1987-2012 with the corresponding HBGA, indicating the presence of a conserved epitope (Lindesmith et al., 2012a, 2014). However, the capacity of the NVB 71.4 mAb to block this interaction differs between the GII.4 strains, suggesting the epitope consists of both conserved and variable

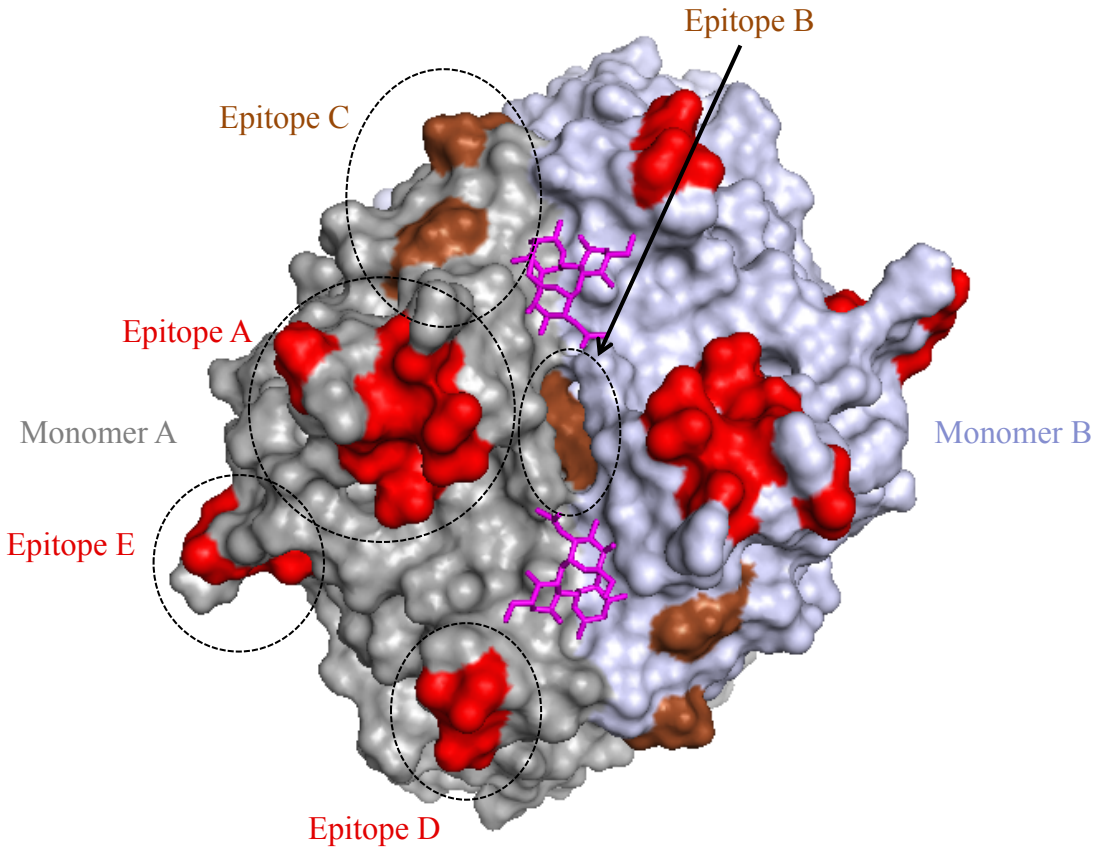


Figure 1.7: Immune epitopes on the surface of the GII.4 capsid. Epitopes are shown on the Sydney 2012 dimeric P domain structure 4WZT. One monomer is shown in grey and the other monomer is shown in blue-white. This is a top view of the capsid. The epitopes are coloured in both monomers but are only labelled in monomer A for clarity. Neutralising epitopes A, D and E are shown in red and putative epitopes B and C are shown in brown. The residues within each epitope are shown by the dashed circle.

residues (Lindesmith et al., 2012a). The exact epitope targeted by NVB 71.4 is yet to be identified, but access to the epitope is conformation and temperature dependent and is regulated by several amino acids, including E316, R484 and K493 (Lindesmith et al., 2014). While sites 316, 484 and 493 are highly conserved within GII.4, site 310 varies in New Orleans 2009 and Sydney 2012 and variation at this site has been demonstrated to alter access to the conserved epitope (Lindesmith et al., 2014).

Several studies have examined the antigenic changes between temporally adjacent pairs of pandemic strains in more detail (Lindesmith et al., 2008, 2012b, 2013; Debbink et al., 2013). While subtle antigenic differences between the Camberwell 1994-like and US95/96 strains have been demonstrated (Lindesmith et al., 2012a), these strains appear to be very similar antigenically, supporting suggestions for a non-immunity driven mech-

anism for the increase in prevalence of the GII.4 capsid in the mid-1990s (Lindesmith et al., 2008, 2011). Epitope E and site 395 within epitope D have been implicated as important in the emergence of Farmington Hills 2002 (Lindesmith et al., 2008, 2012b). Epitope A was demonstrated to change between the Den Haag 2006 and New Orleans 2009 strains, with sites 294, 368 and 372 within epitope A differing between the strains (Lindesmith et al., 2013). Sydney 2012 changed in epitopes A and D compared with its temporal predecessor, New Orleans 2009 (Debbink et al., 2013). Only roughly 30% of the polyclonal blockade response against New Orleans 2009 was retained against Sydney 2012, with a substantial percentage of this reduction being due to changes at sites 294 and, in particular, 368 (Debbink et al., 2013). New Orleans 2009 and Sydney 2012 also differ in epitope D and substitutions at site 373 next to epitope A and site 396 next to epitope D may also have had a role in immune escape (Debbink et al., 2013; Allen et al., 2014). The conserved epitope recognised by the human mAb NVB 71.4 likely also changed between New Orleans 2009 and Sydney 2012 (Debbink et al., 2013; Lindesmith et al., 2014).

Recombination has also been suggested to have a role in the emergence of GII.4 pandemic strains (Eden et al., 2013). Prior to their respective pandemic emergence, New Orleans 2009 acquired a Yerseke 2006-like ORF1, while Sydney 2012 acquired a GII.Pe ORF1 by recombination (Eden et al., 2013). Such recombination events have the potential to enable acquisition of an advantageous ORF1 which could, for example, be able to evade T cell immune responses raised against non-structural proteins (White, 2014). Additional recombination events at the ORF1/ORF2 and ORF2/ORF3 boundaries have been identified leading to epidemic GII.4 strains, leading to suggestions of the importance of recombination in GII.4 evolution (Eden et al., 2013). However, to date no studies have been conducted comparing the properties of non-structural proteins found with the different pandemic GII.4 strains and there is little information available of T cell immunity against these proteins. Therefore, while recombination has clearly occured leading to certain GII.4 strains, the importance of this recombination is unknown. The Sydney 2012 pandemic strain circulated with both the GII.P4 and GII.Pe ORF1s (Wong et al., 2013).

The first four pandemic GII.4 strains were suggested to have evolved from their temporal predecessor, with the exception of Den Haag 2006 which was proposed to have evolved from Farmington Hills 2002 (Siebenga et al., 2007; White, 2014). However, a recent phylogenetic analysis demonstrated that the New Orleans 2009 and Sydney 2012

strains last shared a common ancestor close to the start of the Apeldoorn 2007 epidemic and therefore neither of these strains evolved from a previously circulating pandemic strain (Eden et al., 2014). Pre-pandemic forms of New Orleans 2009 and Sydney 2012 were identified circulating one and two years prior to the onset of the respective pandemic, leading to suggestions that these strains circulated at low level until acquiring substitutions within the capsid P2 domain that drove global emergence (Eden et al., 2014). Several additional studies have now identified individual cases and/or outbreaks caused by GII.4 pandemic strains over several years prior to the onset of their pandemic (Siebenga et al., 2009; Sdiri-Loulizi et al., 2009; Mans et al., 2016; Allen et al., 2016). In a recent study of UK community gastroenteritis cases from the mid 1990s, Allen et al. (2016) identified four of the pandemic strains between eight and 18 years prior to the onset of their respective pandemics and suggest the importance of surveillance of sporadic gastroenteritis in the community (Allen et al., 2016).

Exactly where GII.4 noroviruses may be circulating when they acquire their pandemic enabling mutations has been a topic of active discussion (Karst and Baric, 2015). Immunocompromised patients have been suggested to be a potential reservoir for new GII.4 strains due to the diverse viral population that has been detected within these patients (Bull et al., 2012; Vega et al., 2014b; Karst and Baric, 2015). This diversity is in stark contrast to the limited diversity detected in infections of immunocompetent individuals (Bull et al., 2012; Karst and Baric, 2015) and changes acquired during chronic infection of immunocompromised patients can alter viral antigenicity (Debbink et al., 2014). However, a recent study developed a mathematical model of transmission that incorporates immunocompetent and immunocompromised patients and suggested that immunocompromised patients do not greatly contribute to norovirus evolutionary dynamics due to the rarity and isolation of such hosts (Eden et al., 2017). The potential for elderly or malnourished hosts to act as reservoirs for new strains has also been mooted as these individuals have diminished, but not absent, immune responses (Hickman et al., 2014; Karst and Baric, 2015). Malnourished mice exhibited a more diverse murine norovirus quasispecies, but again there is no evidence that this could result in new variants that are onwardly transmissible (Hickman et al., 2014). Importantly, few studies of norovirus epidemiology have been carried out in developing countries where malnutrition is more common (Mans et al., 2016). GII.4 viruses have been detected in faeces from cows, pigs

and dogs, raising the possibility of zoonotic transmission (Mattison et al., 2007; Summa et al., 2012; Caddy et al., 2015). However, norovirus genotypes typically exhibit strict species tropism and, while inter-species transmissions have clearly occurred in the past, no zoonotic transmission has yet been observed (Karst and Baric, 2015).

1.3.5 Emergence of the GII.P17-GII.17 Kawasaki 2014 lineage in Asia in Winter 2014-2015

In the Winter of 2014-2015, a novel GII.17 lineage replaced GII.4 Sydney 2012 as the dominant cause of norovirus outbreaks in East Asia (Matsushima et al., 2015; De Graaf et al., 2015; Lu et al., 2015; Fu et al., 2015; Lee et al., 2015). Until this time, the GII.17 genotype had only rarely been detected (De Graaf et al., 2015). The novel GII.17 lineage, termed GII.17 Kawasaki 2014, was first detected in wastewater samples in Kenya in 2012 and 2013 (Kiulia et al., 2014; De Graaf et al., 2015) and the origins of this lineage were traced back to Africa, with a single introduction into Asia (Lu et al., 2016). GII.17 Kawasaki 2014 caused a large peak in norovirus outbreaks in Japan, Hong Kong and Jiangsu province, China in Winter 2014-2015 (Matsushima et al., 2015; Fu et al., 2015; Chan et al., 2015), similar to increases associated with the emergence of new pandemic GII.4 strains (Lopman et al., 2004a). The GII.17 Kawasaki 2014 lineage is associated with a novel GII.P17 ORF1 and the surface of the capsid exhibits two insertions and multiple substitutions relative to other GII.17 viruses, including at sites within and close to known epitopes within the GII.4 capsid (Chan et al., 2015; De Graaf et al., 2015; Singh et al., 2016a). It was recently demonstrated that the GII.17 Kawasaki 2014 lineage is antigenically distinct from other GII.17 viruses and sites 393-396 were identified as an important epitope in GII.17 (Lindesmith et al., 2017a). Additionally, GII.17 Kawasaki 2014 was demonstrated to bind to HBGA types A, B and O, suggesting the ability to infect a large proportion of the population (Chan et al., 2015). While sporadic cases of GII.17 Kawasaki 2014 were detected in multiple countries worldwide, this lineage has not replaced GII.4 Sydney 2012 as the dominant norovirus strain in most countries and is therefore largely restricted to East Asia (Chan et al., 2017a).

1.3.6 Emergence of viruses with the GII.P16 ORF1 in Winter 2016-2017

Two novel lineages containing the GII.P16 ORF1 emerged as a major cause of norovirus outbreaks in Winter 2016-2017 (Niendorf et al., 2017; Ao et al., 2017; Lu et al., 2017; Liu et al., 2017; Bidalot et al., 2017). The emergence of these lineages was associated with a large increase in norovirus outbreaks compared with recent years in Germany (Niendorf et al., 2017), China (Ao et al., 2017) and France (Bidalot et al., 2017). While the two lineages each have the GII.P16 ORF1, one lineage has the GII.4 Sydney 2012 capsid, while the other has the rarely detected GII.2 capsid (Matsushima et al., 2016; Niendorf et al., 2017; Chan et al., 2017b). The GII.P16-GII.2 lineage was dominant in Taiwan in late 2016, with outbreaks caused by this lineage predominantly occurring in schools (Liu et al., 2017).

1.4 Human norovirus immunity and immune escape

1.4.1 Innate immunity against noroviruses

Innate immunity has been suggested to be critical for controlling norovirus infection, due to the short duration of symptomatic infections in immunocompetent patients (Karst et al., 2014). Type I interferon (IFN) responses are particularly important for control of MNV infection and mice lacking type I interferon receptor, type I and type II interferon receptors or STAT1 exhibit far higher MNV viral titres compared with wild-type mice and succumb to lethal infection (Karst, 2003; Thackray et al., 2012). In the absence of type I IFN, MNV infection can be controlled by a mechanism involving type II IFN (Karst, 2003; Karst et al., 2014). Additionally, MNV strongly induces a type I IFN response in cultured macrophages and dendritic cells (Thackray et al., 2012) and replicates to higher titres in cultured macrophages and dendritic cells from mice lacking components of the IFN pathway, including $STAT1^{-/-}$, $IRF3^{-/-}$, $MDA5^{-/-}$ and $IFNAR1^{-/-}$ mice, indicating a key role for IFN in control of MNV infection (Wobus et al., 2004; McCartney et al., 2008;

Thackray et al., 2012). However, a single cycle of Norwalk virus replication and replication of a GII.3 virus in human cells able to mount strong IFN responses does not result in an IFN response (Qu et al., 2016). These human noroviruses do not block induction of IFN by co-transfection of Sendai virus, indicating that human noroviruses do not induce IFN (Qu et al., 2016). Additionally, the ability of p48 and p22 to alter protein secretory pathways potentially inhibits the ability of infected cells to mount effective innate immune responses (Roth and Karst, 2016). Interestingly, a single substitution within the MNV p48 homologue NS1/2 is sufficient to enable persistent infection of the colon, consistent with an inhibition of the innate immune response (Nice et al., 2013; Niendorf et al., 2016).

1.4.2 Adaptive immunity against noroviruses

There is evidence of a role for B cells, CD4+ T cells and CD8+ T cells in the control of norovirus infections (Karst et al., 2014). Mice lacking B cells do not clear MNV infections, while the transfer of immune serum or B cells into persistently infected RAG1 knockout mice results in clearance of infection (Chachu et al., 2008b). The presence of long term protective humeral immunity against human noroviruses has been controversial (Debbink et al., 2012b). Volunteer studies have demonstrated the development of short-term immunity against GI.1 norovirus, but also identified that some individuals could be re-infected with the same strain, suggesting infection did not result in long-term protective immunity (Parrino et al., 1977). However, this was challenged by the detection of immune responses in norovirus challenged but uninfected volunteers, leading to suggestions that the high inoculum titres used in early studies were unrealistic and potentially resulted in less protective immunity than would be mounted against a, more realistic, small infectious dose (Lindesmith et al., 2003, 2005, 2012a). An increase in Norwalk virus-specific IgA between one and five days after exposure correlated with protection from infection, in contrast to infected individuals who exhibited an increase in IgA only more than five days after exposure (Lindesmith et al., 2003). The Norwalk virus-specific IgA level had returned to baseline by eight days after exposure, indicating a rapid immunological response and the presence of pre-existing immunity (Lindesmith et al., 2003). The presence of long-term immunity is also supported by the complex binding patterns of human mAbs

against GII.4 strains (Lindesmith et al., 2012a). Mathematical modelling indicates that the duration of human norovirus immunity is 4.1-8.7 years (Simmons et al., 2013). Recently, Parra et al suggested that the GI and GII genogroups could be divided into immunotypes, with each immunotype consisting of one or more closely related genotypes (Parra et al., 2017). Importantly, re-infection within an immunotype is rare, suggesting both protective immunity and antigenic similarity of related genotypes (Parra et al., 2017). Similar to other viruses, the magnitude and duration of norovirus immunity is boosted by multiple exposures, as demonstrated by multiple MNV infections in mice and GI.1 infections of chimpanzees (Chachu et al., 2008a; Bok et al., 2011).

GI- and GII-reactive antibodies are high in acute sera (Reeck et al., 2010; Lindesmith et al., 2010). Cross reactivity patterns differ between the genogroups, with GI sera being cross reactive within the genogroup but not with GII viruses (Debbink et al., 2012b). In contrast, GII sera are far more strain-specific and do not exhibit broad GII-blocking responses (Rockx et al., 2005; Debbink et al., 2012b). Interestingly, there is evidence that even closely related MNV strains can greatly differ in their ability to generate robust immunity (Zhu et al., 2013; Karst et al., 2014). MNV-3 elicits a more robust antiviral antibody and CD4+ T cell response compared with MNV-1, with this robust response resulting in protection from reinfection (Zhu et al., 2013). This difference was due to VP2, with exchange of VP2 between the strains switching the phenotypes, although the mechanism behind this remains unknown (Zhu et al., 2013; Roth and Karst, 2016).

The role of T cells in norovirus infection has not been as well characterised as that of B cells. However, infection or vaccination with human norovirus leads to induction of a primarily CD4+ Th1 cell response, leading to production of IFN-gamma and IL-2 (Lindesmith et al., 2005, 2010). Indeed, an early Th1 response correlated with protection against infection with a GII.2 virus (Lindesmith et al., 2005). T cell responses have been suggested to be cross reactive between closely related genotypes (Lindesmith et al., 2005, 2010). Mice infected with MNV develop an intestinal T cell response by eight days post-infection (Tomov et al., 2013). While mice deficient in CD4+ or CD8+ T cells do clear infection, transfer of immune CD4+ or CD8+ T cells into RAG1 knockout mice reduced chronic viral load and transfer of immune CD4+ T cells into either wild-type of RAG1 knockout mice protects from primary infection (Chachu et al., 2008a; Tomov et al., 2013; Zhu et al., 2013).

1.4.3 Vaccine development against human noroviruses

There is currently no widely accepted treatment for norovirus infection, although compounds targeting various viral and cellular components are under development (Thorne et al., 2016). There has been progress towards the development of a norovirus vaccine, with early clinical trials demonstrating vaccine protection against experimental challenge (Atmar et al., 2011; Bernstein et al., 2015; Nordgren et al., 2016; Cortes-Penfield et al., 2017). Given the high prevalence of norovirus disease, a vaccine eliciting protective immunity in a high proportion of recipients has the potential to provide economic and public health benefits (Bartsch et al., 2012). Several correlates of protection for human norovirus infection have been identified (Ramani et al., 2016) and provide insights into the immunity that an effective vaccine may need to elicit. The presence of blockade antibodies in sera prior to exposure to Norwalk virus was associated with a lower risk of illness in a human challenge study (Reeck et al., 2010). Such antibodies were also associated with protection from infection and disease following vaccination (Atmar et al., 2011, 2015). A key role for IgA in protection from infection and illness is suggested by the results of several studies. Pre-exposure levels of salivary IgA correlate with protection from illness (Ramani et al., 2015), while post-exposure levels correlate with protection from infection as previously discussed (Lindesmith et al., 2003). Additionally, the level of fecal IgA pre-exposure exhibits an inverse relationship with peak virus shedding, while fecal IgA levels seven days post-exposure inversely correlates with the duration of virus shedding (Ramani et al., 2015). A role for IgG is also supported by the presence of virus-specific memory IgG cells correlating with protection from illness (Ramani et al., 2015).

Two vaccines employing different strategies have entered clinical trials and have each demonstrated a degree of protection against experimental infection with the genotype(s) included in the vaccine (Nordgren et al., 2016; Cortes-Penfield et al., 2017). Additional vaccines are in pre-clinical trials. One vaccine in clinical trials is a VLP-based bivalent vaccine containing GI.1 and a consensus GII.4, while the other uses a recombinant adenovirus that expresses the GI.1 capsid (Cortes-Penfield et al., 2017). Lindesmith et al. (2017b) recently showed that while the bivalent vaccine elicited a peak blockade anti-

body avidity against GI.1 35 days after exposure, evidities against heterotypic GI VLPs were not sustained 35 days after vaccination, nor after the same time period following exposure in volunteer studies. As only secretor-positive individuals retined high avidity blockade antibodies 180 days post-exposure, host genetics can influence vaccine response (Lindesmith et al., 2017b). Interestingly, avidity to the consensus GII.4 VLP peaked earlier at day seven and remained elevated through to day 180. Unlike GI.1, this avidity was not dependent on secretor status.

However, the high genetic and antigenic diversity of human noroviruses presents a serious challenge to vaccine development. As previously discussed, genogroup-specifc immunity has been suggested to be elicited by natural infection, with little cross-reactivity between GI and GII noroviruses (Malm et al., 2014, 2015). Additionally, available evidence indicates that newly emerging pandemic GII.4 strains are capable of evading immunity raised against previous pandemic strains (Lindesmith et al., 2012b, 2013; Debbink et al., 2013). However, a recent study suggested that serum from many individuals possesses cross-reactive blockade antibodies that can recognise GII.4 strains years before the strain emerges (Sharma et al., 2017). Avidity against the GII.4 Sydney 2012 strain was induced by the bivalent GI.1 and consensus GII.4 vaccine, despite this strain not emerging until one year after the vaccine trial (Lindesmith et al., 2017b). Therefore this vaccine has the potential to protect individuals against future GII.4 strains.

As previously discussed, norovirus infection rates are particularly high in children under the age of five (O'Brien et al., 2016). It has been suggested that immunisations that elicit robust immunity in adults may not be equally efficacious in children (Cortes-Penfield et al., 2017), raising a potential challenge in vaccine development. Additionally, immunocompromised patients and the elderly are at higher risk of morbidity and mortality from norovirus infection. Both of these groups exhibit differing degrees of immune system impairment that may limit vaccine efficacy. Finally, as previously discussed, norovirus immunity may be short lived. A vaccine would need to elicit long lasting immunity to be cost-effective (Cortes-Penfield et al., 2017).

1.4.4 Thesis aims and organisation

The emergence of pandemic GII.4 strains has been well documented and multiple studies have suggested immune evasion as a mechanism enabling the emergence of these strains. However, the sources of new pandemic strains and the process by which these strains emerge remain poorly understood. In this thesis, we aim to answer: Where do new pandemic GII.4 strains come from? How do new GII.4 strains emerge within the population and spread pandemically? Which genetic changes enable new pandemic strains to emerge? Why do new GII.4 pandemics occur? In chapter 2, we examine the mechanisms that drove the emergence of GII.4 as the dominant genotype in the mid-1990s. In chapter 3, we reconstruct the temporal evolutionary history of the GII.4 genotype to investigate the sources of subsequent pandemic GII.4 strains and suggest mechanisms responsible for norovirus pandemics. We next reconstruct the spatiotemporal history of individual norovirus strains in chapter 4 to investigate their early spread and characterise global circulation. In this chapter, we also identify capsid substitutions that may have been important for the pandemic emergence of the five most recent pandemic strains. Finally, in chapter 5, we demonstrate the worldwide circulation of a newly emerging norovirus lineage and identify the genetic substitutions that may be responsible for the increased circulation of this lineage.

Chapter 2

The GII.4 lineage became pandemic in the mid-1990s due to substitutions in the capsid and/or VP2

2.1 Abstract

Despite the large diversity of norovirus genotypes infecting humans, the GII.4 capsid genotype causes the majority of cases and outbreaks and has caused six pandemics since the mid-1990s. Current evidence suggests that prior to the first pandemic, GII.4 was far less prevalent. While several hypotheses have been suggested for the high prevalence of GII.4, no previous study has examined the viral genetic changes that occured within this genotype as it transitioned from low frequency to high frequency. Here, we carry out phylogenetic analyses on each region of the viral genome and suggest that changes within the capsid and/or VP2 likely drove the increase in GII.4 prevalence. The substitutions leading to the common ancestor of the pandemic GII.4 viruses suggest two mechanisms by which this increase could have been driven: an increase in capsid stability and/or an increase in the range of HBGA-binding. The ORF1 regions found with pandemic GII.4 capsids last shared a common ancestor in the 1970s and this region is therefore very unlikely to have driven pandemic emergence in the mid-1990s. We find evidence for an increase in substitution rate in the RdRps associated with the GII.4 capsid, consistent with an increase in mutation rate. However, this is shared with other RdRp genotypes not found with the GII.4 capsid and was likely acquired by 1906, indicating that an increased

mutation rate was not the driver of pandemic emergence. We demonstrate that GII.4 is the only GII capsid genotype that exhibits an accumulation of change at the amino acid level, consistent with previous results that non-GII.4 genotypes remain 'static' through time. This accumulation of amino acid change was already occurring within GII.4 prior to pandemic emergence. We therefore hypothesise that the GII.4 capsid was pre-adapted for pandemic emergence due to an inherent ability to accommodate amino acid change and the associated RdRps enabling rapid mutation to aid transmission and evasion of host immunity. The increase in capsid stability and/or increase in HBGA-binding range then enabled more efficient transmission and/or infection of a larger proportion of the population, respectively, which drove the actual increase in prevalence of GII.4.

2.2 Introduction

While the GII.4 genotype currently accounts for the majority of human norovirus cases and outbreaks (Kroneman et al., 2008; Siebenga et al., 2009), this high prevalence likely began with the onset of the first GII.4 pandemic in the mid-1990s (Donaldson et al., 2008; Siebenga et al., 2010). To date there have only been limited studies on norovirus molecular epidemiology from prior to the onset of this first pandemic. However, each of the studies including surveillance undertaken prior to 1995 have suggested GII.4 to be one of a number of low prevalence genotypes (Boon et al., 2011; Siqueira et al., 2017; Mori et al., 2017b,a). These studies have been undertaken in both children and adults and within community, hospital and foodborne outbreak settings, suggesting a universally lower prevalence.

Previous studies have resulted in several hypotheses for the high prevalence of the GII.4 genotype. HBGA-binding studies have demonstrated the US95/96 GII.4 pandemic strain can bind to a wider range of HBGAs than other genotypes (Lindesmith et al., 2008; Donaldson et al., 2010). Specifically, US95/96 can bind to HBGAs expressed by secretor positive individuals with A, B and O blood groups and is therefore thought to be able to infect roughly 80% of the human population, a greater proportion than other genotypes (Lindesmith et al., 2008). However, this wide range of HBGA-binding has not been demonstrated for all GII.4 strains, as would be expected if this were the major determi-

nant of high prevalence. For example, the Hunter 2004 pandemic strain does not bind efficiently to any tested synthetic HBGAs under experimental conditions (Lindesmith et al., 2008; Donaldson et al., 2010; Lindesmith et al., 2012a).

The GII.P4 RdRp associated with the majority of pandemic GII.4 strains exhibits a higher mutation rate than the lower prevalence GII.Pb and GII.P7 RdRps (Bull et al., 2010). This has led to suggestions that the GII.4 capsid may be highly prevalent due to the associated GII.P4 RdRp enabling faster change and thereby rapid evasion of existing host immunity (Bull et al., 2010; Bull and White, 2011). Alternatively, a higher mutation rate may increase transmissibility, as recently demonstrated for murine norovirus (Arias et al., 2016). Boon et al demonstrated that while the GII.4 capsid has accumulated amino acid change through time, the GII.3 capsid has remained relatively constant at the amino acid level (Boon et al., 2011). It is therefore possible that the GII.4 capsid is able to accomodate increased amino acid variation compared to the other genotypes, while maintaining important capsid functions (Boon et al., 2011). This is supported by a recent study that demonstrated a wide range of non-GII.4 genotypes have remained static at the amino acid level through time (Parra et al., 2017). However, whether this accumulation of amino acid change within the GII.4 genotype is a cause or consequence of its high prevalence is yet to be determined. The large size of the genogroup GII capsid compared with capsids in the GI genogroup has been suggested be of importance for the ability of GII to accomodate change (Donaldson et al., 2010). In particular, the presence of large loops on the surface of the capsid may enable accomodation of rapid change.

Previous studies have not examined the viral genetic changes that occurred leading to the highly prevalent GII.4 viruses. Here, we carry out phylogenetic analyses and demonstrate that substitutions within either the capsid and/or VP2 likely drove the increase in GII.4 prevalence in the mid-1990s. We identify two potential mechanisms based on the substitutions within the capsid region: increased capsid stability and expansion of the HBGA-binding range. We demonstrate that the common ancestor of pandemic ORF1 sequences occurred in the 1970s, indicating that ORF1 was highly unlikely to have driven pandemic emergence. Phylogenetic data do support a higher substitution rate within the RdRps associated with the GII.4 capsid, consistent with an increased mutation rate, although this is not unique to the RdRps found with the GII.4 capsid and was likely acquired by the early 1900s. Therefore the increase in GII.4 prevalence was not driven by

the acquisition of a higher mutation rate. While most GII capsid genotypes exhibit an accumulation of nucleotide change through time, GII.4 is the only GII genotype to exhibit accumulation of amino acid change. This accumulation of amino acid change was already evident in the pre-pandemic GII.4 lineages. We conclude that the GII.4 capsid was pre-adapted to become pandemic due to its ability to accomodate amino acid change and the high mutation rate of the associated RdRps, and an increase in capsid stability and/or increase in HBGA-binding range were the final steps that enabled the pandemic emergence and subsequent dominance of this genotype.

2.3 Materials and Methods

2.3.1 GII.4 capsid dataset assembly and phylogenetic analyses

To obtain a dataset of all available norovirus GII.4 capsid sequences, we searched GenBank with the term 'norovirus' and sequence length 400-10000 nucleotides. We genotyped each of these sequences using the norovirus genotyping tool (Kroneman et al., 2011) and retained all sequences with the GII.4 capsid that contained more than 800 nucleotides in the capsid region. We removed sequences that within their GenBank record are associated with a patent or a laboratory host and sequences that were listed as a synthetic construct, a laboratory strain, a passaged virus, a cloned sample or an unverified sequence. Any of these associations may have resulted in changes to the sequence after the isolation of the virus from an infected patient and can therefore mislead evolutionary analyses. We also removed sequences that exhibited a frameshift within the capsid region and sequences that contained a stop codon within each frame in the capsid region as these changes will likely have resulted in a non-functional capsid. We removed sequences for which a collection date could not be obtained. We additionally removed one sequence that was listed as having a non-functional capsid. Where multiple longitudinal sequences from the same patient are present, only the earliest collected sequence from the patient was retained.

The root-to-tip distance accumulated by each sequence in a phylogenetic tree rela-

tive to its date of collection gives a measure of the amount of change accumulated per unit time (Rambaut et al., 2016). Sequences that are overly divergent (i.e. exhibit a large root-to-tip distance relative to their collection date) are often spurious sequences, potentially due to sequencing error. Any such sequences should therefore be removed from further phylogenetic analyses. We aligned the GII.4 capsid dataset at the amino acid level using MUSCLE (Edgar, 2008) and reconstructed a nucleotide maximum likelihood tree using RAxML v8.1 (Stamatakis, 2014) with the GTR substitution model and gamma rate heterogeneity with four gamma classes. We assessed topological robustness using 1000 bootstrap replicates. We examined the temporal signal within this tree using TempEst v1.5 (Rambaut et al., 2016) and did not find any sequences that were overly divergent based on their collection date. This phylogenetic tree was therefore used for further analyses. The final GII.4 capsid dataset contained 2198 capsid sequences and included sequences from all of the major GII.4 strains (Table 2.1).

To calculate the date of the common ancestor of the pandemic GII.4 clade and the date of the increase in prevalence of the GII.4 capsid, we extracted the CHDC 1970s, Tokyo 1980s, Bristol 1993, Camberwell 1994-like, Kaiso 2003, US95/96 and Osaka 2007 sequences from the GII.4 dataset (Table 2.1). This dataset contains all of the available sequences from the early part of the GII.4 phylogenetic tree, with it being well supported that Osaka 2007 is the first strain that diverges after US95/96. There is a correlation between root-to-tip distance and sampling date within this dataset, as calculated using a nucleotide maximum likelihood tree reconstructed with RAxML as above. We reconstructed the temporal evolutionary history of the early GII.4 lineage using BEAST v.2.4.2. We used the SRD09 substitution model, where the alignment is partitioned into codon positions 1 and 2 and codon position 3, with a separate HKY substitution model with gamma rate heterogeneity and four gamma classes being applied to each partition. We used the relaxed lognormal clock model and applied a lognormal prior to the substitution rate with mean 6.83×10^{-3} and standard deviation 0.1. This prior was chosen based on the mean and 95% HPD posterior estimates of the GII.4 substitution rate estimated in chapter 3 (Table 3.2). We applied a coalescent Bayesian skyline tree prior. Three replicate runs were carried out and run until convergence, as assessed using Tracer v1.5. The replicate runs were combined with removal of suitable burnin using LogCombiner v2.2.1 and the maximum clade credibility (MCC) phylogenetic tree was identified using TreeAnnotator v2.2.1.

GII.4 strain	Number of capsid sequences	Number of VP2 sequences
CHDC1970s	8	8
Tokyo 1980s	2	2
Bristol 1993	3	3
Camberwell 1994-like	7	5
US95/96	79	5
Farmington Hills 2002	92	61
Lanzhou 2002	11	2
Asia 2003	42	16
Kaiso 2003	5	0
Hunter2004	59	17
Den Haag 2006	858	629
Yerseke 2006	46	18
Apeldoorn 2007	69	36
Osaka 2007	30	4
New Orleans 2009	396	158
Sydney 2012	480	157
GII.4 could not assign strain	22	6

Table 2.1: Summary of the GII.4 capsid and VP2 datasets. All sequences were genotyped using the norovirus genotyping tool (Kroneman et al., 2011). There is currently no genotyping system for VP2 and each VP2 sequence was therefore assigned to a strain on the basis of its capsid strain, as assigned using the norovirus genotyping tool (Kroneman et al., 2011). Phylogenetic analysis confirmed that the majority of sequences clustered similarly in the capsid and VP2, with the exception of Osaka 2007 where the 11 sequences with the Osaka 2007 capsid clustered within two separate clades within the VP2 phylogenetic tree. Four of the sequences cluster with the US95/96 VP2, while the other seven sequences cluster within the Den Haag 2006 clade. We classified the four sequences that cluster with US95/96 as having the Osaka 2007 VP2, as this is consistent with the clustering of the RdRp classified as Osaka 2007. The other seven sequences were included in the Den Haag 2006 clade. Pandemic strains are labelled in bold font.

We calculated the date at which the pandemic GII.4 clade diverged from the pre-pandemic GII.4 lineages from the date of the node immediately upstream of the common ancestor of the pandemic GII.4 clade in all trees in the posterior distribution. We calculated the date of the increase in GII.4 prevalence using information from the Bayesian skyline plot in each sampled step in the MCMC chain. The Bayesian skyline plot is a piecewise constant model to estimate relative genetic diversity (estimated by the effective population size multiplied by generation time, $Ne\tau$) through time. Here, $Ne\tau$ can change between time windows but is constant within a window. We employed a Bayesian skyline plot consisting of five time windows. The value of $Ne\tau$ within each window and the length of each window are integrated over during the MCMC chain. We initially calculated the date at which each sampled step supported an increase in $Ne\tau$ of more than 100% relative to that in the first time window. Most sampled steps supported an increase in the 1990s, while a small number of steps supported an earlier increase between the 1960s and 1980s (Figure S2.1). Examination of the steps supporting an early increase demonstrated that these steps supported a small early increase in $Ne\tau$ followed by another, larger, increase in $Ne\tau$ in the mid-1990s (Figure S2.1). We therefore defined the time of pandemic onset in these steps as the time at which $Ne\tau$ increased by more than 400% relative to baseline. We combined the date at which these steps supporting a small early increase in $Ne\tau$ increased by more than 400% relative to baseline with the time at which the remaining steps support an increase in $Ne\tau$ of more than 100% relative to baseline to obtain a single distribution for the time of the increase in GII.4 capsid prevalence (Figure S2.1). As this distribution is not normally distributed (see Figure 2.2 panel D), we used the median as the point estimate of the date of the increase in frequency. The date at which the median $Ne\tau$ begins to increase in frequency in the Bayesian skyline plot reconstructed on all sampled steps using Tracer v1.5 (October 1994) is very similar to our estimate of the median date of the increase in GII.4 prevalence (December 1994).

We inferred the nonsynonymous substitutions that occurred leading to the pandemic GII.4 clade common ancestor using the nucleotide maximum likelihood tree reconstructed on the 2198 sequence GII.4 capsid dataset and ten bootstrap tree topologies reconstructed during bootstrapping of this dataset. We used multiple tree topologies to assess the robustness of the inferred substitutions to tree topology. We used RAxML v8.1 (Stamatakis, 2014) to optimise branch lengths to the amino acid alignment using the translated GII.4

alignment and the WAG substitution model with optimised base frequencies. We carried out ancestral reconstruction at the amino acid level with PAML v4.9 (Yang, 2007) using the WAG substitution matrix and optimised base frequencies. We identified nonsynonymous substitutions by comparing the amino acid sequence of the pandemic GII.4 clade common ancestor with that of the upstream node. Each of the nonsynonymous substitutions inferred to have occurred leading to the pandemic GII.4 clade common ancestor occurred with each tested tree topology, with the exception of T534A which occurred in 9 of 11 tested tree topologies. We verified the identified changes using SubRecon (available at https://github.com/chrismonit/SubRecon) which calculates the probability of each nonsynonymous substitution occurring at each site along a branch of interest. We ran SubRecon on the maximum likelihood tree topology with WAG-optimised branch lengths and employed the WAG substitution matrix incorporating the base frequencies optimised by RAxML. The same nonsynonymous substitutions were identified using SubRecon as through ancestral reconstruction with PAML. SubRecon identified the probability of each substitution to be 1.0 with the exception of T534A, the probability of which was 0.86. The conservation of sites of interest was assessed using the coloured trees method, where each tip within the nucleotide maximum likelihood phylogenetic tree was coloured by the amino acid residue within that sequence at the site.

We identified sites under different selective constraints in the pandemic GII.4 clade compared with the pre-pandemic GII.4 lineages using TdG v1.1.2 (Tamuri et al., 2009). Briefly, this method identifies amino acid sites that are evolving differently within two or more subclades. A substitution matrix is optimised for each subclade independently at each site and a likelihood ratio test performed to determine whether the two subclade-specific rate matrices fit the data significantly better than a single rate matrix optimised on the entire dataset. Two matrices fitting the data significantly better than one is indicative of the two clades evolving differently at the site. We used the maximum likelihood tree reconstructed on the 2198 GII.4 sequence dataset with WAG-optimised branch lengths and optimised base frequencies. We labelled the viruses from the CHDC 1970s, Tokyo 1980s, Bristol 1993, Camberwell 1994-like and Kaiso 2003 strains as pre-pandemic and the viruses within the pandemic GII.4 clade were labelled as pandemic.

We carried out homology modelling using two methods: SWISS-MODEL (Biasini et al., 2014) and I-TASSER (Zhang, 2008). We constructed homology models of the

common ancestor of the pandemic GII.4 clade and the immediately upstream ancestor using the sequences of these nodes reconstructed by PAML. We modelled the structure of the P domain excluding the last nine residues because currently available GII.4 structures consist of this region. These ancestral sequences therefore differ at sites 285, 294, 309, 333, 340, 395, 459, 497 and 505. We identified the best fitting template structure using SWISS-MODEL, which in each case was a dimeric US95/96 P domain structure in complex with HBGA type B (Cao et al., 2007), PDB identifier 2OBT. We used structure 2OBT as the template for SWISS-MODEL. However, I-TASSER does not model the structure of dimeric complexes and instead models monomeric structures. While structure 2OBT was solved as a dimer, the PDB structure file only contains a single monomeric capsid. We therefore used structure 5IYQ as the template for homology modelling with I-TASSER, which is a CHDC 1970s structure in complex with HBGA type A. Here, prior to modelling, we altered the PDB structure file for this sequence to convert the two chains of the dimer into a single chain, with all atoms retaining the same spatial position. Analysis of the resulting structures was carried out using PyMol v1.74.

We investigated the hydrogen bond network formed by site 459 in the capsid using PyMol. We examined interactions within two hydrogen bonds of site 459 in nine solved crystal structures and four homology models. We used six P domain crystal structures from three pandemic GII.4 strains: Farmington Hills 2002 not bound to a HBGA (PDB identifier 400V), Farmington Hills 2002 bound to Lewis antigen B (4OPS), Den Haag 2006 bound to Lewis antigen A (4WZL), Den Haag 2006 bound to HBGA type B (4X06), Sydney 2012 bound to HBGA type A (4WZT) and Sydney 2012 bound to HBGA type B (4OP7). Each of these six structures has glutamine at site 459, enabling comparison of the interaction network formed by this residue in different structures from different pandemic strains. We additionally examined the interaction network in three P domain structures from a pre-pandemic GII.4 virus in the CHDC 1970s strain, either not bound to a HBGA (PDB identifier 5IYN), bound to HBGA type A (5IYP) or bound to HBGA type B (5IYQ). These structures have serine at site 459. We also investigated the interaction network in the homology models of the pandemic GII.4 clade common ancestor and the immediately upstream ancestor constructed using SWISS-MODEL and I-TASSER. The common ancestor of the pandemic GII.4 clade has glutamine at site 459 while the immediately upstream ancestor has leucine at site 459.

2.3.2 RdRp and ORF1 dataset assembly and phylogenetic analyses

We collected a dataset containing all of the 1248 available norovirus RdRp sequences using the same approach as in collation of the capsid dataset, but in this case retained sequences from all genogroups (Table 2.2). We aligned the RdRp dataset at the amino acid level using MUSCLE (Edgar, 2008) and screened for recombination using the single breakpoint (SBP) method in HyPhy (Kosakovsky Pond et al., 2005; Pond et al., 2006). Briefly, this method splits the alignment at each variable site and reconstructs a phylogenetic tree on either side of the breakpoint. The fit of these two phylogenetic trees to the two alignments is compared with the fit of a single phylogenetic tree for the whole alignment using the Akaike information criteria (AIC). A significantly better fit of the two trees to the split alignment is taken as evidence of recombination, with one or more sequences having a different evolutionary history on either side of the breakpoint. No recombination was identified in the RdRp dataset. We therefore reconstructed a nucleotide maximum likelihood tree using RAxML (Stamatakis, 2014) as above. To analyse the accumulation of nucleotide change through time, we midpoint rooted the maximum likelihood tree, which results in the GV genogroup diverging from the root. We obtain the same root location when rooting to maximise the correlation between root-to-tip distance and collection date with TempEst v1.5 (Rambaut et al., 2016), although there is no appreciable correlation between these values. We calculated the total distance from the root of the tree to each tip in nucleotide substitutions/site and divided by the collection date of the tip to obtain a measure of the accumulation of nucleotide change per unit time. The observed division in the nucleotide change per unit time within the GII genogroup was also identified in a maximum likelihood nucleotide phylogenetic tree reconstructed on all GII RdRp sequences and rooted on a single GIV.1 outgroup sequence (accession number of outgroup JQ613567.1).

We calculated the substitution rate within the RdRp of the GII.P4 lineage and the rest of the GII clade independently using BEAST v2.4.2 (Bouckaert et al., 2014). Due to the large number of sequences in our dataset from several strains (Table 2.2), we took five random sequences from the Farmington Hills 2002, Hunter 2004, Den Haag 2006,

Genotype	Number of sequences in ORF1 dataset	Number of sequences in RdRp dataset
GII.P1	6	6
GII.P2	1	1
GII.P3	5	5
GII.P4	852	836
GII.P5	1	1
GII.P6	3	3
GII.P7	42	40
GII.P8	2	2
GII.P11	1	1
GII.P12	36	27
GII.P13	1	1
GII.P15	1	1
GII.P16	9	9
GII.P17	56	55
GII.P18	1	1
GII.P20	1	1
GII.P21	29	29
GII.P22	17	17
GII.Pa	2	2
GII.Pc	3	3
GII.Pe	85	85
GII.Pg	17	16
GII.Pj	1	1
GII.Pm	2	2
GII.Pp	1	1
GII could not assign genotype	4	4
GI	NA	28
GIII	NA	7
GIV	4	4
GV	NA	57
GVI	NA	2

Table 2.2: Summary of ORF1 and RdRp datasets. We assembled datasets containing all available norovirus RdRp or ORF1 sequences and genotyped each sequence using the norovirus genotyping tool (Kroneman et al., 2011). The number of sequences included from each genotype is shown for the ORF1 dataset and the RdRp dataset. We only retained sequences in the GII or GIV genogroups in the ORF1 dataset.

New Orleans 2009, Yerseke 2006 and Apeldoorn 2007 GII.P4 strains and five random sequences from the GII.Pe clade associated with the GII.4 Sydney 2012 capsid. We used the SRD09 model of nucleotide substitution and a relaxed lognormal clock model. We employed a diffuse prior on the substitution rate with mean $6.73x10^{-3}$ and standard deviation 0.79, chosen to encompass our estimate of the substitution rate within the RdRps found with the GII.4 capsid in chapter 3 (Table 3.2). The results were highly similar when employing an alternative lognormal prior on the substitution rate with mean $3x10^{-3}$ and standard deviation 0.79. We applied a coalescent Bayesian skyline tree prior. Three replicate runs were carried out for each dataset and were run until convergence, as assessed with Tracer v1.5. We combined the replicate runs after removal of suitable burnin using LogCombiner v2.2.1 and identified the MCC tree using TreeAnnotator v2.2.1.

We calculated the distribution of the substitution rate parameter in the GII.P4 lineage and the rest of the GII clade using the posterior estimates of the mean and standard deviation of the lognormal distribution of substitution rates within each clade. We used a two sample Kolmogorov Smirnov test to calculate the statistical significance of the difference between these lognormal distributions.

To test for a difference in the substitution rate at the third codon position in each clade, we ran BEAST on the third codon position only. We used the HKY substitution model with amongst site rate variation accomodated by a gamma distribution with four gamma categories. We used a lognormal clock model and employed a lognormal prior distribution with mean $6.673x10^{-3}$ and standard deviation 0.79 for the mean substitution rate. A difference in subsitution rate was tested for using the two sample Kolmogorov Smirnov test as above.

To identify the nonsynonymous substitutions that occurred within the RdRp leading to the GII.P4 lineage, we initially optimised branch lengths within the RdRp maximum likelihood phylogenetic tree with RAxML (Stamatakis, 2014) using the translated RdRp alignment, the WAG substitution matrix and optimised base frequencies. Ancestral reconstruction was carried out at the amino acid level using PAML v4.9 (Yang, 2007) as above and nonsynonymous substitutions that occurred leading to the GII.P4 lineage were identified by comparing the sequence at the common ancestor of the GII.P4 lineage with that at the upstream node. We identified sites under different selective constraints in the GII.P4 lineage and the rest of the GII clade using TdG (Tamuri et al., 2009) as described

above. With TdG, the GII.P4 lineage was labelled as one clade and the rest of the GII clade was labelled as the second clade.

The divergence date between the GII.Pe-Sydney 2012 RdRps and the pandemic GII.P4 RdRps was calculated using the posterior distribution of trees for the GII.P4 lineage BEAST runs. We identified the date of the most recent common ancestor of these sequences in each tree in the posterior distribution and calculated the mean and 95% HPD of this distribution.

We reconstructed the nonsynonymous substitutions that occurred leading to the pandemic GII.P4 and GII.Pe clades using a dataset containing all 1183 GII and GIV ORF1 sequences available (Table 2.2), collated as described for the capsid dataset above. We included the GIV sequences to enable accurate rooting of the GII clade. We aligned the ORF1 dataset at the amino acid level using MUSCLE (Edgar, 2008). However, the 5' end of P48 and a region within P22 do not align well between the more distantly related genotypes within GII or between GII and GIV. It was therefore not possible to accurately identify homologous sites within these regions across all of the genotypes. Mis-alignment would result in the inclusion of non-homologous sites, which may mislead ancestral reconstruction. At sites which could not be accurately aligned across all genotypes, we retained the site in GII.P4, GII.Pe and all other genotypes that aligned well with GII.P4 and GII.Pe, but removed the site from genotypes that could not accurately be aligned with GII.P4 and GII.Pe. As expected, the genotypes that cluster more closely with GII.P4 and GII.Pe in the ORF1 phylogenetic tree could be aligned within these regions, while the more distantly related genotypes could not be aligned. Therefore at certain sites within P48 and P22, the genotypes more distantly related to GII.P4 and GII.Pe had gap characters. As the genotypes closely related to GII.P4 and GII.Pe did not have residues deleted, we would expect the nonsynonymous substitutions reconstructed close to GII.P4 and GII.Pe to remain reliable. Genotypes GIV.1, GIV.2, GII.P6, GII.P7, GII.P8, GII.P11, GII.P15, GII.P18, GII.P20 and GII.P22 had residues deleted at the 5' end of P48 and within P22. Genotype GII.Pp had residues deleted at the 5' end of P48 but could be aligned within P22 and so did not have residues deleted in this region.

We found no evidence of recombination within the ORF1 dataset as screened for using SBP. We reconstructed a maximum likelihood phylogenetic tree of the ORF1 dataset using RAxML (Stamatakis, 2014) and optimised branch lengths within this tree using the

translated ORF1 alignment as described above. We carried out ancestral reconstruction at the amino acid level as above and identified nonsynonymous substitutions close to GII.P4 and GII.Pe by comparing the reconstructed sequences at the relative ancestal nodes. To ensure the identified substitutions were robust to tree topology, we carried out the same analysis on ten bootstrap tree topologies in addition to the maximum likelihood tree.

2.3.3 GII.4 VP2 dataset assembly and phylogenetic analyses

We collected a dataset containing all of the available VP2 sequences from viruses with a GII capsid as described in the collation of the capsid dataset above (Table 2.1). There is currently no genotyping system for VP2, with genotyping instead being carried out on the RdRp and capsid (Zheng et al., 2006; Kroneman et al., 2013). While recombination has been noted close to the boundary between ORFs 2 and 3 (Eden et al., 2013), to our knowledge there are no reported instances of inter-genogroup recombination in this region. It is therefore very likely that viruses with a GII capsid will also be found with a VP2 that can be classified as GII based on phylogenetic clustering. We also included four VP2 sequences where the ORF1 and capsid regions were not sequenced. We reconstructed a nucleotide maximum likelihood phylogenetic tree of this 1387 sequence dataset using RAxML (Stamatakis, 2014) as described above, which confirmed that the VP2 sequences found with the GII.4 capsid form a well supported monophyletic clade. Two of the viruses where the ORF1 and capsid were not sequenced were identified within the GII.4 VP2 clade: sequence EF635440.1 clusters with Asia 2003 and sequence EU876893.1 clusters with Yerseke 2006.

We inferred the nonsynonymous substitutions leading to the pandemic GII.4 clade using the GII nucleotide maximum likelihood VP2 tree. We optimised branch lengths within this tree and carried out ancestral reconstruction as described above. Nonsynonymous substitutions were identified by comparing the sequence at the common ancestor of the pandemic GII.4 clade with the sequence at the immediately upstream node.

To reconstruct the evolutionary dynamics of the early part of the GII.4 clade, we used the sequences that cluster within the CHDC 1970s, Bristol 1993, Camberwell 1994-like, US95/96, Farmington Hills 2002 and Osaka 2007 clades. The VP2 sequences from

viruses with the Osaka 2007 capsid cluster within two regions of the GII.4 clade, with four sequences clustering with US95/96 and seven sequences clustering with Den Haag 2006. Here, we included only the four Osaka 2007 sequences that cluster with US95/96. We initially reconstructed a maximum likelihood phylogenetic tree of this dataset using RAxML (Stamatakis, 2014) and confirmed the dataset exhibited accumulation of nucleotide change through time using TempEst v1.5 (Rambaut et al., 2016). We used BEAST v2.4.2 (Bouckaert et al., 2014) to reconstruct the temporal evolutionary history of the early GII.4 VP2 lineage, using the SRD09 nucleotide substitution model and a relaxed lognormal clock model. We employed a lognormal prior distribution on the mean substitution rate with mean 6.11×10^{-3} and standard deviation 0.1, chosen to encompass the mean and 95% HPD posterior estimates of the GII.4 VP2 substitution rate estimated in chapter 3 (Table 3.2).

2.3.4 Comparison of accumulation of change between GII.4 and other GII genotypes

We collected datasets for each GII capsid genotype containing all of the available capsid sequences from that genotype. The GII.1, GII.2, GII.3, GII.5, GII.6, GII.7, GII.12, GII.13, GII.14, GII.17 and GII.21 contained sufficient sequences for further analysis (Table 2.3), in addition to the GII.4 dataset assembled previously. Sequences were aligned within each dataset at the amino acid level using MUSCLE (Edgar, 2008). We screened each of the genotype datasets for recombination using SBP (Kosakovsky Pond et al., 2005; Pond et al., 2006). In cases where a breakpoint was found ($p < 0.05$ in the Kishino-Hasegawa (KH) test), we reconstructed a nucleotide maximum likelihood tree on either side of the proposed breakpoint using RAxML (Stamatakis, 2014) as described above and identified sequences that clustered differently on either side of the breakpoint with strong bootstrap support. These sequences were considered to be putative recombinants and were removed from further analyses. Potentially recombinant sequences were identified in the GII.12, GII.14 and GII.21 datasets (Table 2.3); the sequences removed are listed in Table S2.1. As our aim was to remove potentially recombinant sequences rather than to identify recombination, we did not carry out any further analyses to confirm these

Genotype	Number of sequences	Date span of sequences	Putative recombinant sequences
GII.1	22	1971-2014	No
GII.2	103	1975-2011	No
GII.3	134	1975-2014	No
GII.5	23	1978-2013	No
GII.6	63	1975-2014	No
GII.6 clade 1	21	1976-2012	NA
GII.6 clade 2	21	1975-2013	NA
GII.6 clade 3	21	1977-2014	NA
GII.7	15	1976-2010	No
GII.12	28[a]	1990-2010	Yes
GII.13	47	1983-2013	No
GII.14	16[a]	1978-2013	Yes
GII.17	178	1978-2015	No
GII.21	19[a]	2005-2015	Yes

Table 2.3: Summary of genotype datasets. The number and time span of sequences within each genotype dataset is shown. Genotypes were screened for recombination using SBP, putative recombinant sequences are listed in Table S2.1. NA - recombination was not tested for within the GII.6 subclades as recombination was not present within the complete GII.6 dataset. [a] The number of sequences following removal of putatively recombinants is shown for GII.12, GII.14 and GII.21.

recombination events.

We reconstructed a nucleotide maximum likelihood tree for each genotype using RAxML (Stamatakis, 2014) as described above. The temporal signal for each genotype was investigated using TempEst v1.5 (Rambaut et al., 2016). We identified the best fitting root location by minimising the heuristic residual mean squared score for the correlation between root-to-tip distance and sampling date. We divided the GII.6 genotype into three subclades due to a lack of temporal signal within the complete genotype dataset and reconstructed a nucleotide maximum likelihood tree on each subclade independently.

To calculate the accumulation of amino acid change through time, we used the tree reconstructed on the complete capsid nucleotide dataset. We only carried out this analysis for genotypes that exhibited a correlation between nucleotide root-to-tip distance and collection date. We optimised branch lengths within the tree with RAxML using the translated capsid alignment and the WAG substitution matrix with optimised base frequencies. We rooted the tree on the branch where it was rooted to maximise the correlation between nucleotide root-to-tip distance and collection date. As the root could have occurred at any

point along this branch, we rooted at ten equally spaced intervals along this branch and used the root location that maximised the R^2 correlation between root-to-tip distance and collection date. To compare the accumulation of amino acid change within the pandemic GII.4 clade and the pre-pandemic GII.4 lineages, we extracted the respective clade from the GII.4 nucleotide tree and optimised branch lengths independently using the respective amino acid alignments and the WAG substitution model with optimised base frequencies. We plotted the amino acid root-to-tip distance versus collection date for each sequence and calculated the best fit regression line usig TempEst. Both datasets exhibit a correlation between amino acid root-to-tip distance and collection date.

2.4 Results

2.4.1 The first GII.4 pandemic likely began in the mid-1990s

We reconstructed a phylogenetic tree containing 2198 GII.4 capsid sequences from all of the major GII.4 strains (Figure 2.1). All of the GII.4 strains associated with pandemics or large epidemics since the mid-1990s form a well-supported monophyletic clade downstream of the US95/96 common ancestor (starred node in Figure 2.1). We therefore defined the sequences downstream of the US95/96 common ancestor as forming the pandemic GII.4 clade and those lineages that branch prior to the US95/96 common ancestor as the pre-pandemic GII.4 lineages. The sequences in the pre-pandemic lineages belong to the CHDC 1970s, Tokyo 1980s, Bristol 1993, Camberwell 1994-like and Kaiso 2003 strains (Figure 2.1). The only sequences that branch prior to the US95/96 common ancestor that were collected after 1994 are those belonging to the Kaiso 2003 strain, with Kaiso 2003 likely evolving from the Bristol 1993 strain (Figure 2.1). The common ancestor of the pandemic GII.4 clade occurred in July 1992 (95% HPD December 1990-August 1993) and the lineage leading to the pandemic GII.4 clade diverged from other sampled GII.4 capsid sequences in April 1987 (95% HPD December 1983-September 1989) (Figure 2.2 panel A). The GII.4 capsid most likely increased in prevalence in late 1994 (95% HPD January 1994-October 1996), although there is significant support on the increase occur-

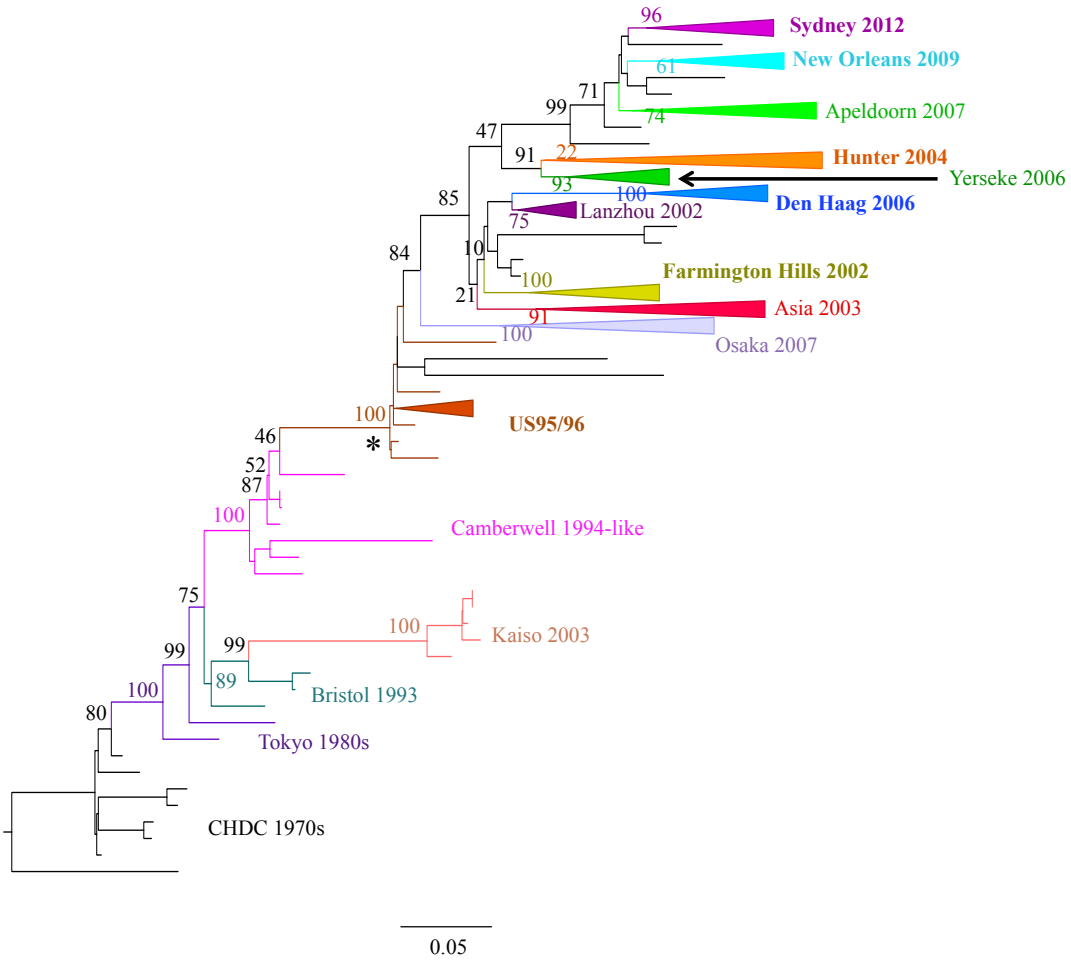


Figure 2.1: Maximum likelihood phylogenetic tree of the GII.4 capsid. A nucleotide maximum likelihood tree was reconstructed for 2198 GII.4 capsid sequences using RAxML. Major GII.4 strains are collapsed for clarity. Branches are coloured by strain to match the colour of the strain label. Pandemic strains are labelled in bold font. The starred node is the common ancestor of the pandemic GII.4 clade. Bootstrap supports are shown on trunk nodes, with the colour of the bootstrap support at strain root nodes matching the colour of the strain label. The scale bar shows the expected number of nucleotide substitutions per site.

ring earlier in 1994 and in either 1995 or 1996 (Figure 2.2 panels C and D). However, there is very little support on an increase in relative genetic diversity prior to 1994, supporting previous epidemiological and phylodynamic data suggesting that US95/96 was the first GII.4 pandemic and coincided with increased prevalence of the GII.4 capsid (Siebenga et al., 2010; Donaldson et al., 2008).

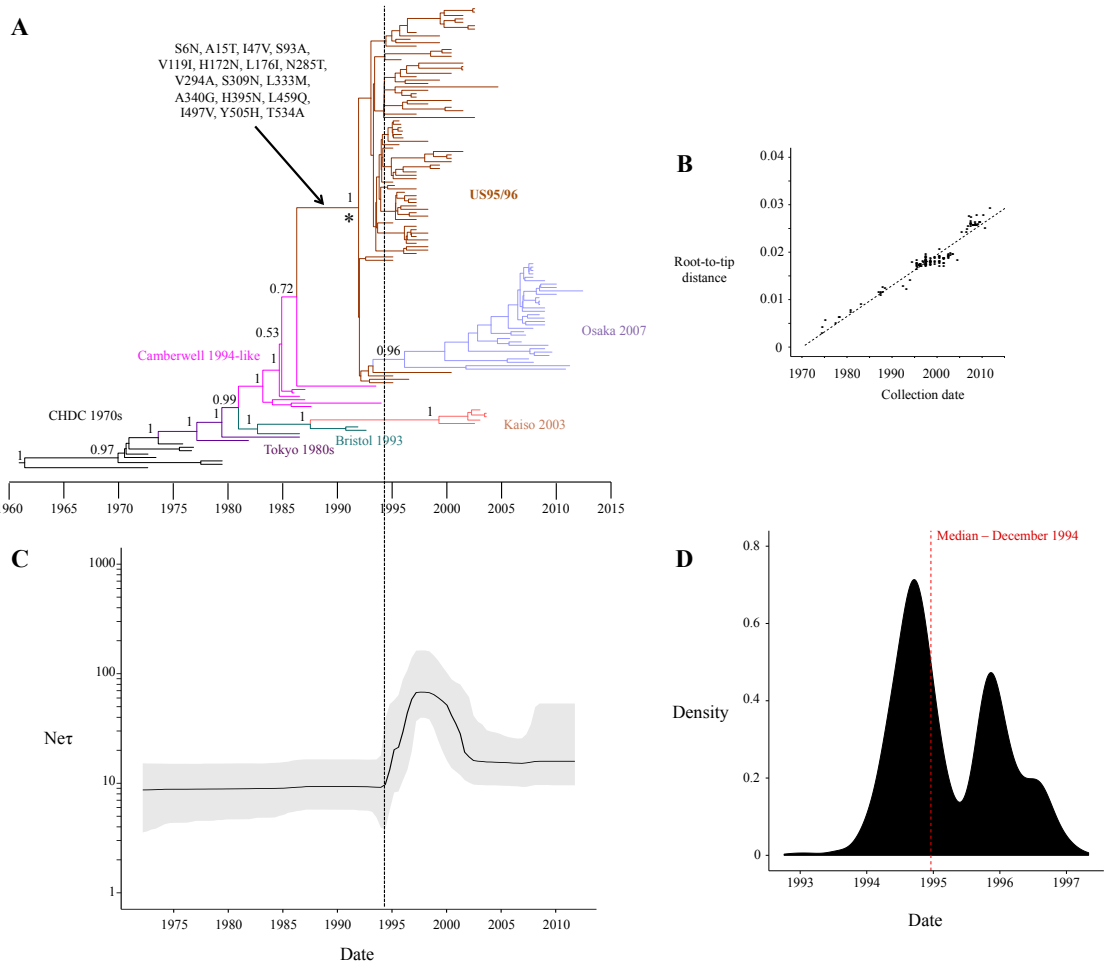


Figure 2.2: Temporal evolutionary history of the early GII.4 lineage. (**A**) Maximum clade credibility tree of the early sequences from the GII.4 lineage: CHDC 1970s, Tokyo 1980s, Bristol 1993, Camberwell 1994-like, Kaiso 2003, US95/96 and Osaka 2007. The starred node is the common ancestor of the pandemic GII.4 clade. The nonsynonymous substitutions that occurred leading to the common ancestor of the pandemic clade are labelled. Branches are coloured by the corresponding strain and match the colour of the strain label. Posterior supports are shown at nodes along the trunk of the tree. The pandemic US95/96 strain is labelled in bold font. (**B**) The correlation between root-to-tip distance and collection date within the early GII.4 lineage, calculated from a nucleotide maximum likelihood tree reconstructed using the same sequences in the MCC tree in panel A. The strong correlation between root-to-tip distance and collection date enabled reconstruction of the temporal evolutionary history of the lineage. (**C**) Bayesian skyline plot of the early GII.4 lineage, plotting a measure of the relative genetic diversity (Neτ) through time. The increase in relative genetic diversity indicates an increase in prevalence. The dashed line represents the start of the increase in relative genetic diversity in October 1994. (**D**) The distribution of the date of onset of the first GII.4 pandemic, estimated using the date of the first increase in Neτ at each sampled step in the MCMC chain. The red dashed line represents the median date of the increase in frequency. All values within the 95% HPD of the date of pandemic onset are within the range shown; there are a small number of steps that support an earlier increase in frequency; these steps are not shown here for clarity but are shown in Figure S2.1.

2.4.2 Capsid substitutions that may have been important for the pandemic emergence of the GII.4 genotype

We would expect that capsid sites important for the increase in prevalence of GII.4 in the mid-1990s would have undergone a substituion leading to the common ancestor of the pandemic GII.4 clade and would exhibit different amino acid residues in the pandemic GII.4 clade and the pre-pandemic GII.4 sequences. We infer that 17 nonsynonymous substitutions occurred leading to the pandemic GII.4 clade common ancestor: S6N, A15T, I47V, S93A, V119I, H172N, L176I, N285T, V294A, S309N, L333M, A340G, H395N, L459Q, I497V, Y505H and T534A (Figures 2.2 panel A, 2.3). Of these, sites 6, 15, 47, 93, 119, 172 and 176 are located within the shell domain, sites 459, 497, 505 and 534 are located within the P1 domain and sites 285, 294, 309, 333, 340 and 395 are located within the P2 domain (Figure 2.3). The substitutions at sites 6, 15, 47, 119, 309 and 534 are unlikely to have been important for the increase in prevalence of the GII.4 capsid as these sites exhibit the same amino acid residues in the pandemic GII.4 clade and the pre-pandemic GII.4 sequences (Figure 2.3). Using TdG (Tamuri et al., 2009), we identify two capsid sites that are evolving under different selective constraints in the pandemic GII.4 clade compared with the pre-pandemic GII.4 viruses with a false discovery rate < 0.05: sites 333 and 459 (Figure 2.3). Site 333 is located within the P2 domain, while site 459 is in the P1 domain.

The two sites that were identified as evolving under different selective constraints in the pandemic GII.4 clade versus the pre-pandemic GII.4 viruses (sites 333 and 459) are both located within the dimerisation interface. Four regions of inter-monomer interactions have been identified within the P domain (Cao et al., 2007). Site 333 is located within region I, in which the capsid monomers interact via hydrophobic interactions (Cao et al., 2007). This site is conserved as leucine within the pre-pandemic GII.4 viruses and changes regularly between methionine and valine within the pandemic GII.4 lineage (Figure 2.4). While leucine is intermediate between methionine and valine in size and hydrophobocity (Kyte and Doolittle, 1982), it has been suggested that valine and methionine exhibit a greater propensity to form beta sheet secondary structure than leucine (Koehl and Levitt, 1999). Site 333 is located close to the center of beta strand $\beta 4$ within

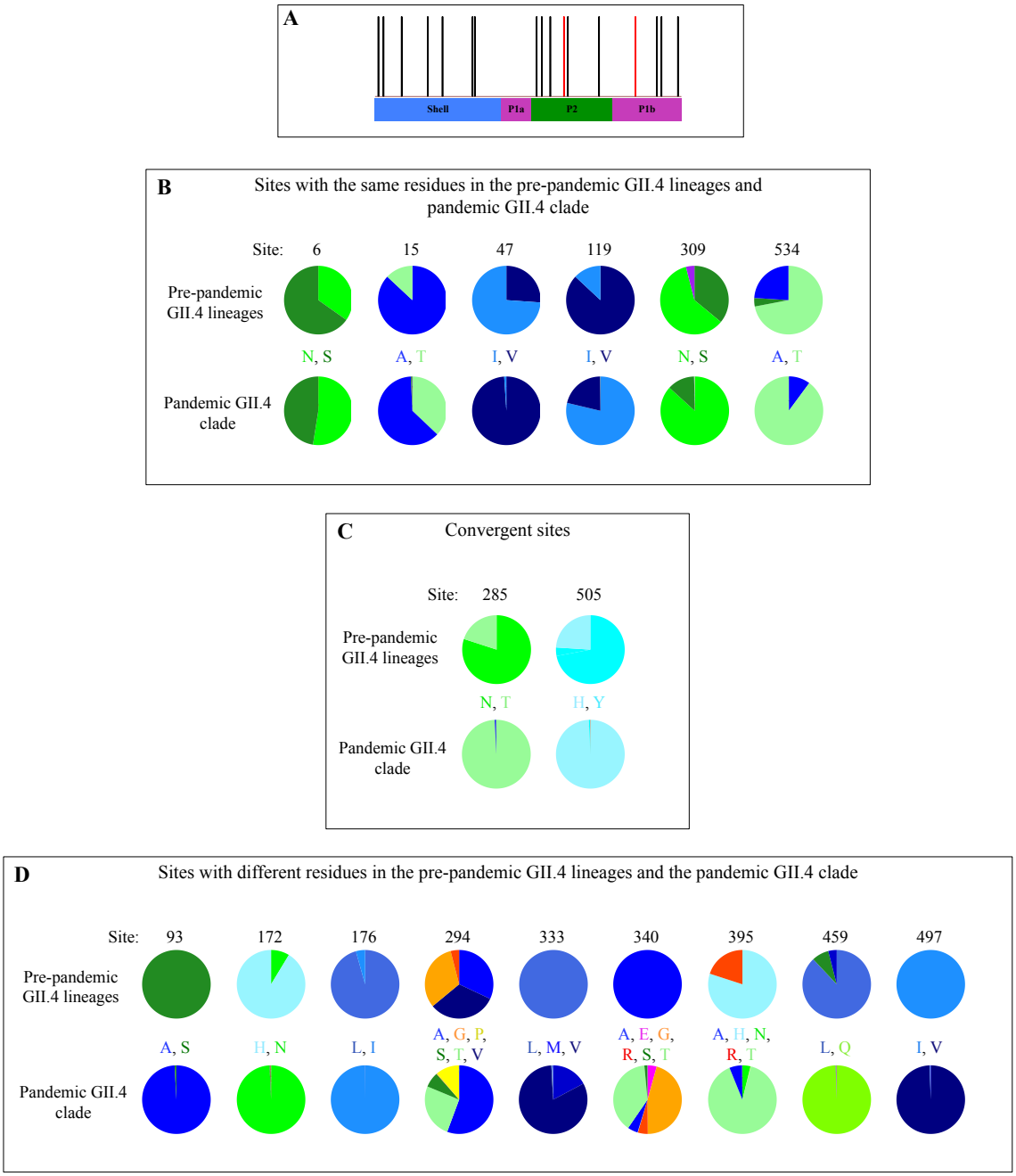


Figure 2.3: Nonsynonymous capsid substitutions leading to the pandemic GII.4 clade. (**A**) Location within the capsid of the 17 nonsynonymous substitutions that occurred leading to the common ancestor of the pandemic GII.4 clade. Sites 333 and 459 that change leading to the common ancestor of the pandemic GII.4 clade and are under different selective constraints in the pandemic GII.4 clade and pre-pandemic GII.4 lineages are shown in red. The sites that change leading to the common ancestor of the pandemic GII.4 clade and were not inferred to be under different selective constraints are shown in black. (**B**) Amino acid residue distribution in the pandemic GII.4 clade versus the pre-pandemic GII.4 lineages at each site inferred to have undergone a substitution leading to the pandemic GII.4 clade but exhibits the same amino acid residues in the pre-pandemic GII.4 lineages and the pandemic GII.4 clade. Amino acid residues at each site are shown between the respective pie charts and residues with similar properties are shown in similar colours. (**C**) As in panel B but sites that exhibit a convergent substitution leading to the pandemic GII.4 clade and the Kaiso 2003 epidemic strain are shown. (**D**) As in panel B but sites that underwent a substitution leading to the pandemic GII.4 clade and exhibit a different amino acid residues in the pre-pandemic GII.4 lineages and pandemic GII.4 clade are shown.

the P2 domain beta barrel structure.

Site 459 is located within region IV of the dimerisation interface and this region contains a hydrophilic center formed by the side chains of multiple residues (Cao et al., 2007). Site 459 is largely conserved as leucine within the pre-pandemic GII.4 lineages and is conserved as glutamine within the pandemic GII.4 clade (Figure 2.5 panel A). In solved P domain crystal structures from pandemic GII.4 strains, glutamine 459 forms a hydrogen bond network through its main chain and side chain atoms (Figures 2.5, S2.2). It is unfortunately not possible to directly compare the hydrogen bond networks formed by glutamine 459 and leucine 459 as the only pre-pandemic GII.4 virus with a solved P domain crystal structure has serine at site 459, a rare residue at this site within GII.4. However, the hydrogen bond network formed by glutamine 459 in pandemic GII.4 viruses is more extensive than that formed by serine 459 in a pre-pandemic GII.4 virus (Figures 2.5, S2.2). To enable comparison of the interaction network formed by glutamine 459 and leucine 459, we constructed P domain homology models of the common ancestor of the pandemic GII.4 clade (glutamine at site 459) and the immediately upstream node (leucine at site 459). Homology models support glutamine 459 forming an increased number of hydrogen bonds compared with leucine 459, due to side chain interactions formed by glutamine 459 that are not possible with the hydrophobic leucine 459 side chain (Figure 2.5). However, the exact interaction network differs between homology models constructed with different methods. Together, the solved crystal structures and homology models indicate the possibility for an increased number of inter-monomer contacts with glutamine at site 459 compared with the ancestral leucine residue at this site.

It has previously been suggested that an expansion in the range of bound HBGAs may have been responsible for the increase in GII.4 prevalence (Lindesmith et al., 2008; Donaldson et al., 2008, 2010). Site 395 mutated leading to the common ancestor of the pandemic GII.4 clade and is located within a flexible loop that undergoes a structural rearrangement to allow binding to certain HBGAs (Singh et al., 2015). While site 395 has not been demonstrated to directly bind to HBGAs (Singh et al., 2015), previous studies have demonstrated that mutations at this site can influence HBGA binding, in particular binding to HBGA type A (Lindesmith et al., 2008). Site 395 is conserved as histidine within the pre-pandemic GII.4 lineages, with the exception of the Kaiso 2003 strain which has arginine at this position (Figure 2.6). Site 395 mutated from histidine to asparagine

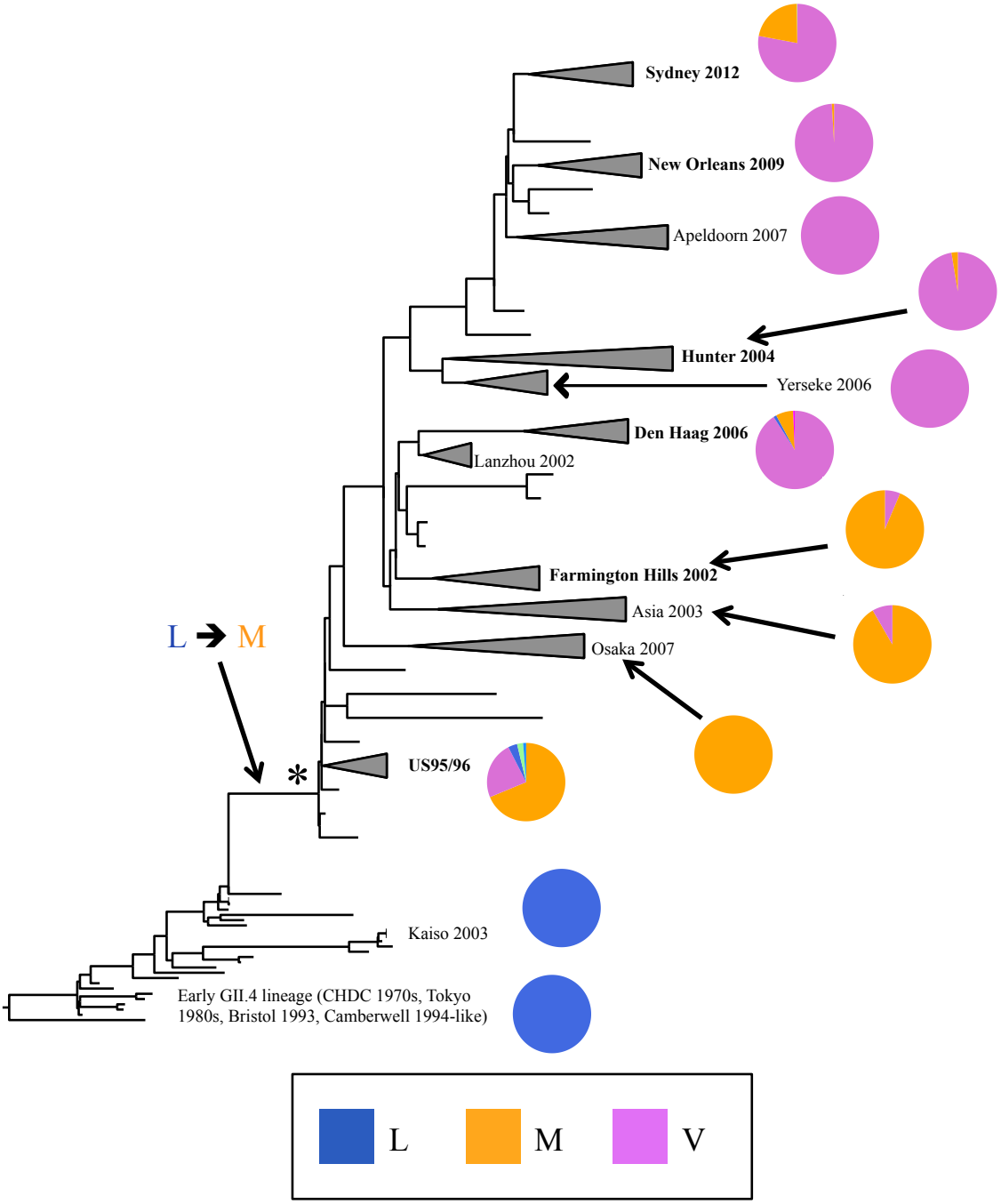


Figure 2.4: Variation at capsid site 333. Shown is the distribution of amino acid residues at site 333 throughout the GII.4 genotype. The early GII.4 lineage (consisting of the CHDC 1970s, Tokyo 1980s, Bristol 1993 and Camberwell 1994-like strains) and Kaiso 2003 strain have leucine at this site. The site underwent a substitution from leucine to methionine leading to the common ancestor of the pandemic GII.4 clade (the starred node) and was then variable between methionine and valine within the pandemic GII.4 clade. The distribution of residues is shown for each major strain within the pandemic GII.4 clade and demonstrates that multiple substitutions occurred between methionine and valine at this site. The tree shown is the same as that in Figure 2.1.

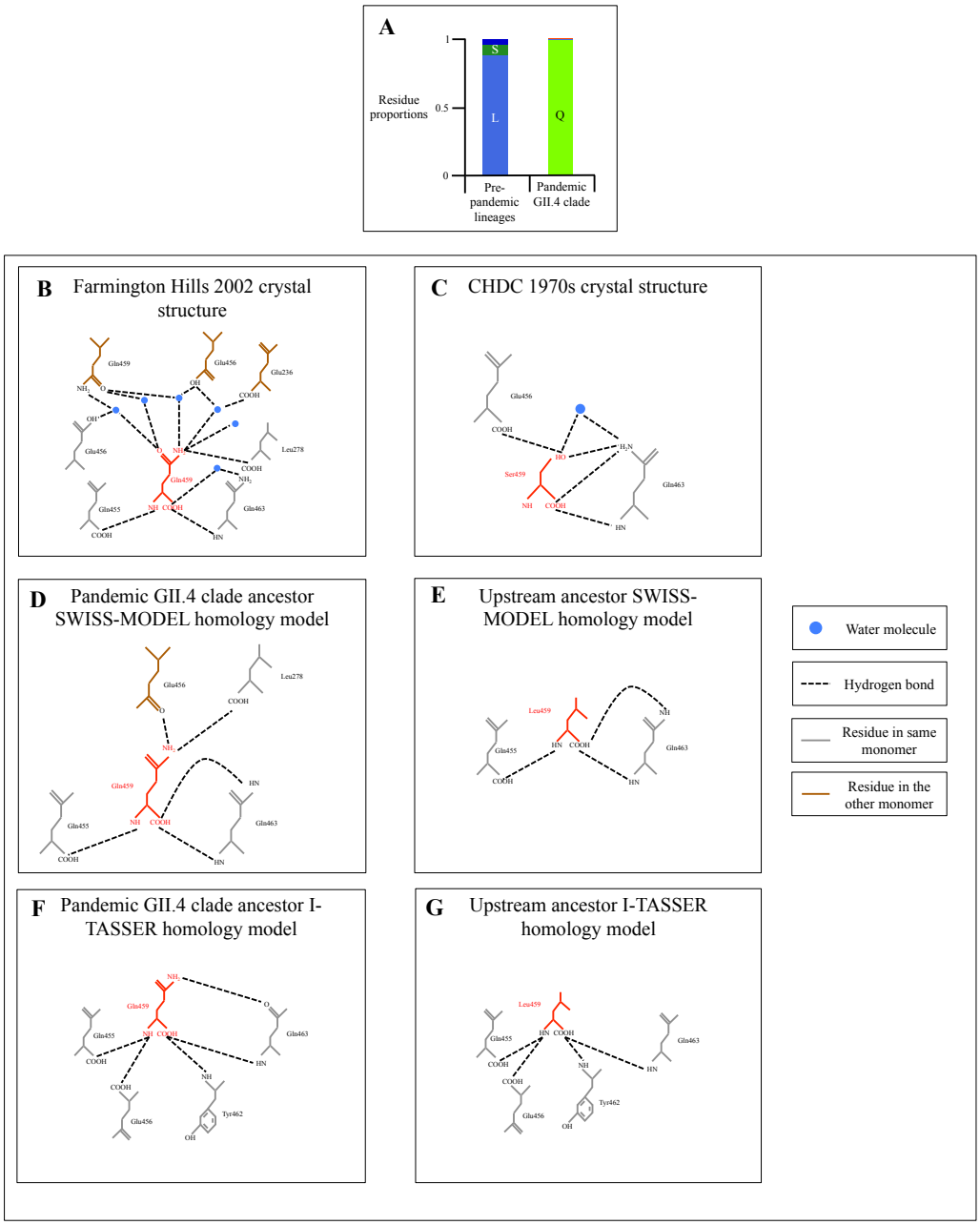


Figure 2.5: The interaction network at site 459. **(A)** The distribution of amino acid residues at site 459 within the pre-pandemic GII.4 lineages and the pandemic GII.4 clade. **(B-G)** The interaction network formed by site 459 in solved crystal structures (**B,C**) and homology models (**D-G**). The residue at site 459 is shown in red in each case and residues that hydrogen bond with site 459 either directly or through a single water molecule intermediate are shown in grey if the residue is in the same monomer or brown if the residue is in the other monomer. Water molecules are shown as blue circles and each hydrogen bond is shown as a black dashed line. (**B**) The interaction network formed by glutamine 459 in a solved Farmington Hills 2002 crystal structure, PDB identifier 400V. The interaction network formed by glutamine 459 is compared between six solved crystal structures from the pandemic GII.4 clade in Figure S2.2. (**C**) The interaction network formed by serine 459 in a solved CHDC 1970s crystal structure, PDB identifier 5IYP. The interaction network formed by serine 459 is compared between three solved crystal structures from a CHDC 1970s virus in Figure S2.2. (**D-E**) Homology models of the pandemic GII.4 clade common ancestor (**D**) and the immediately upstream ancestor (**E**) as determined by SWISS-MODEL. (**F-G**) Homology models of the pandemic GII.4 clade common ancestor (**F**) and the immediately upstream ancestor (**G**) as determined by I-TASSER.

leading to the pandemic GII.4 clade and was conserved as asparagine in the US95/96 strain (Figure 2.6). However, the Hunter 2004, Den Haag 2006, New Orleans 2009 and Sydney 2012 pandemic strains have threonine conserved at this site and site 395 is largely conserved as threonine in Farmington Hills 2002. Of the epidemic strains, Yerseke 2006 and Osaka 2007 have threonine, Apeldoorn 2007 has alanine and Asia 2003 is variable between alanine and threonine at site 395 (Figure 2.6). Interestingly, it was mutation of alanine to threonine that was previously demonstrated to alter binding to HBGA type A (Lindesmith et al., 2008). The residues at site 395 in the pandemic GII.4 clade (alanine, asparagine and threonine) are all smaller than the histidine residue present in the pre-pandemic GII.4 lineages. Given the importance of movement of the 391-395 loop for HBGA-binding, it is possible that a decrease in residue size at site 395 may have resulted in a more flexible loop, enabling more efficient binding to a greater range of HBGAs. Alternatively, both residues found within the pre-pandemic GII.4 viruses are positively charged and the loss of the positive charge at this position may have had the same effect. To test this hypothesis, we suggest future experiments to validate whether VLPs with alanine, asparagine or threonine at site 395 exhibit broader HBGA-binding compared to VLPs with histidine at this site. In particular, we would test the HBGA-binding profile of the common ancestor of the pandemic GII.4 clade and upon mutation of site 395 to A, T and H within this ancestral sequence.

Of the sites that change leading to the common ancestor of the pandemic GII.4 clade, site 294 is located within blockade epitope A, site 333 is located within putative epitope B, site 340 is located in putative epitope C and site 395 is located within blockade epitope D (Lindesmith et al., 2012a). Sites 294 and 340 are highly variable within the pandemic GII.4 clade and site 294 is also highly variable within the pre-pandemic GII.4 sequences (Figures 2.3, 2.7).

The potential importance of substitutions at sites 93, 172, 176, 285, 497 and 505 is less clear, as these sites are not located within a part of the capsid with known function. However, each of these sites is highly conserved within the pandemic GII.4 clade, with the only substitutions occurring either in individual tip viruses or within small clades where the viruses within the clade were sampled over a very short time period (Figure 2.7). Therefore there is no evidence that viruses within the pandemic GII.4 clade that mutate at any of these six sites can persist and transmit within the population. It is therefore

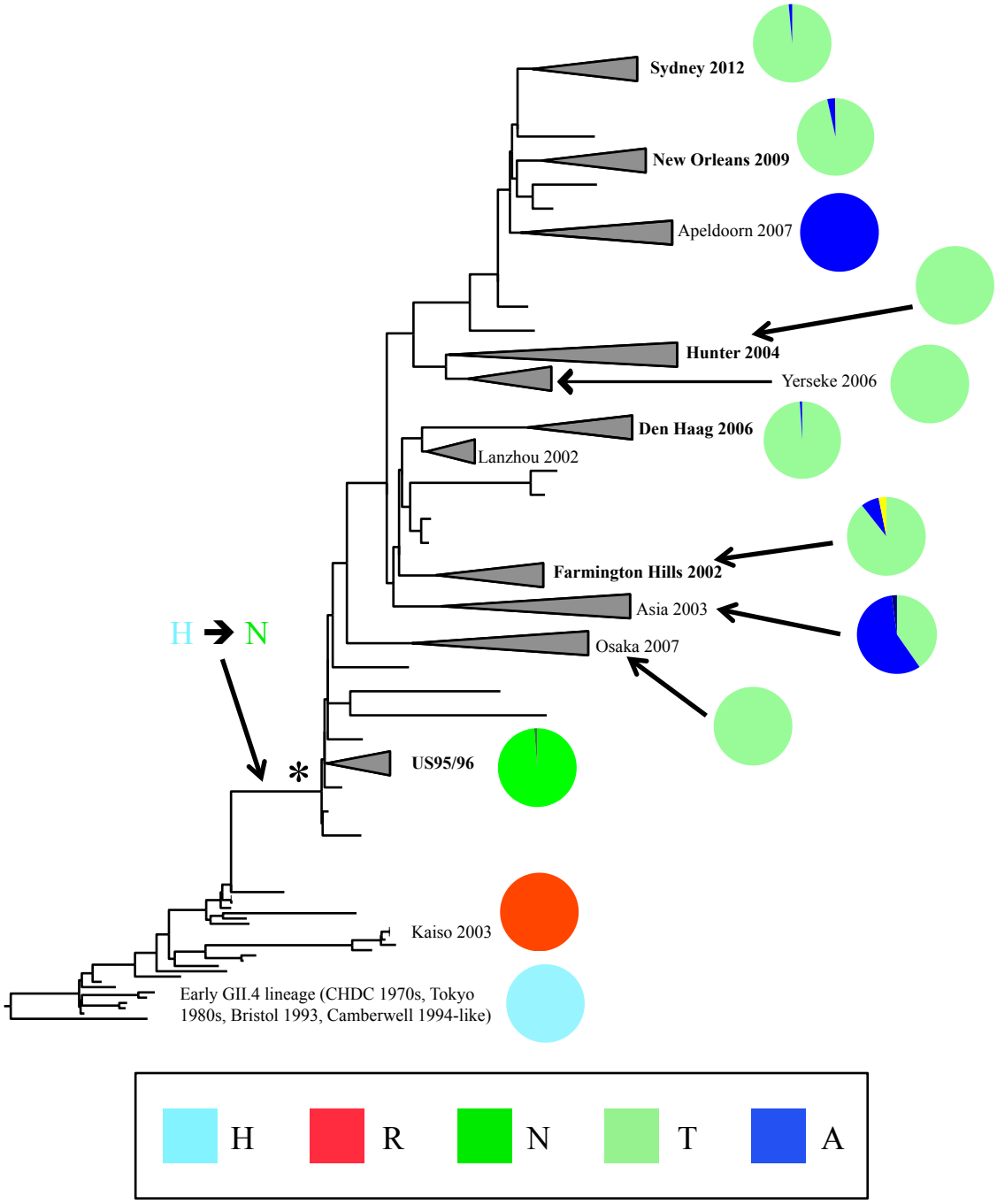


Figure 2.6: Substitution at site 395 within the HBGA-binding loop. Shown is the distribution of amino acid residues at site 395 throughout the GII.4 genotype. The early GII.4 lineage (consisting of the CHDC 1970s, Tokyo 1980s, Bristol 1993 and Camberwell 1994-like strains) has histidine at this site, while the Kaiso 2003 strain has arginine. The site underwent a substitution from histidine to asparagine leading to the common ancestor of the pandemic GII.4 clade (the starred node) and was then variable between alanine, asparagine and threonine within the pandemic GII.4 clade. The distribution of residues is shown for each major strain within the pandemic GII.4 clade. The tree shown is the same as that in Figure 2.1.

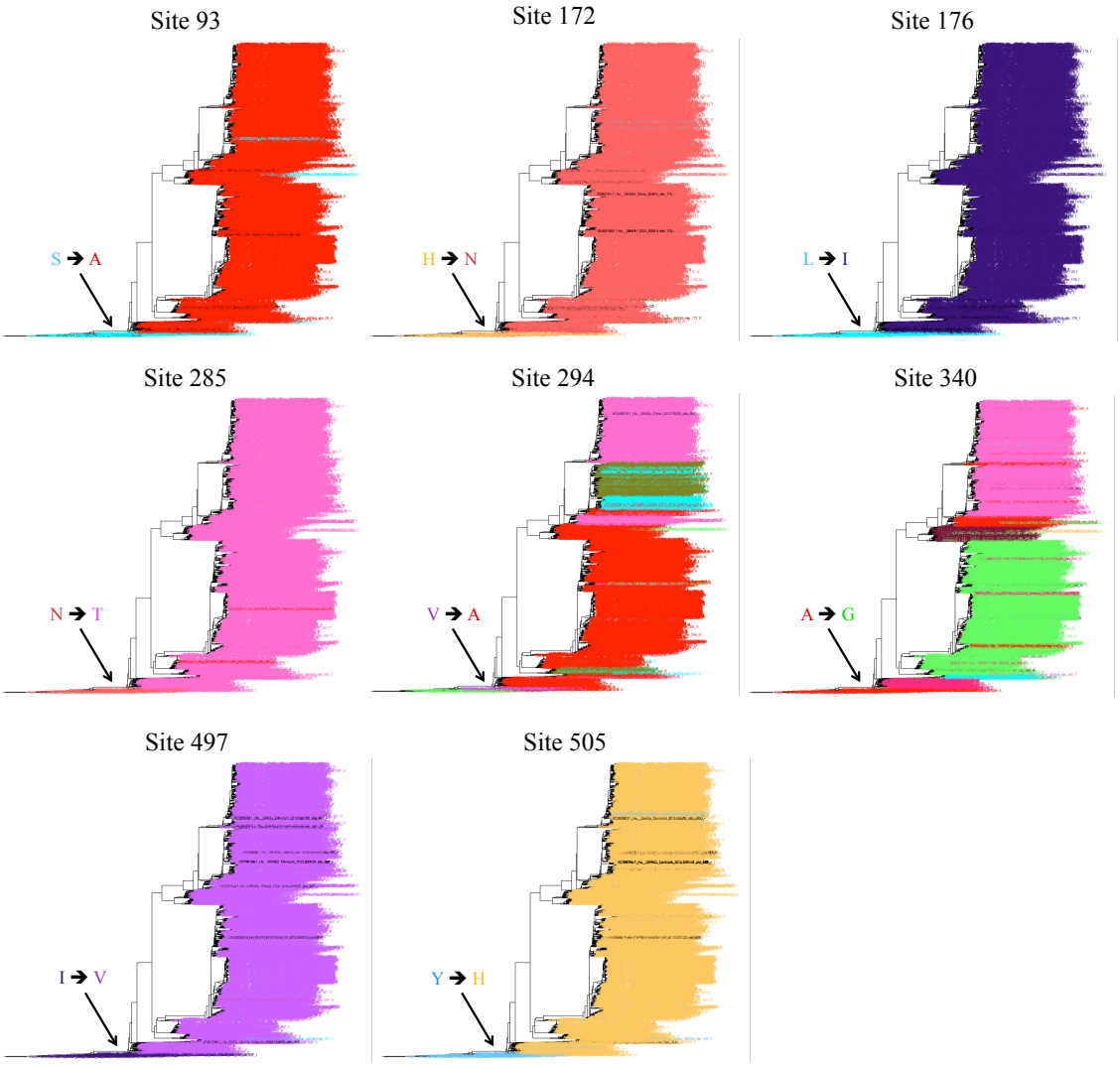


Figure 2.7: Conservation of sites within the pandemic GII.4 clade. Coloured trees for sites 93, 172, 176, 285, 294, 340, 497 and 505. The substitution at each site leading to the pandemic GII.4 clade is labelled. Sites 93, 172, 176, 285, 497 and 505 are largely conserved within the pandemic GII.4 clade, while sites 294 and 340 are highly variable. Black indicates that the site was not sequenced within that sample.

possible that either these sites have important functions within the pandemic GII.4 clade and/or that these sites cannot easily mutate away from the consensus residue without destabilising the capsid structure. Sites 285 and 505 exhibit convergent changes along the branches leading to the pandemic GII.4 clade and the Kaiso 2003 common ancestor (Figure S2.3), indicating that the substitutions at these sites are not sufficient to enable pandemic spread.

2.4.3 The RdRps found with pandemic GII.4 capsids last shared a common ancestor in the 1970s

The first five pandemic GII.4 strains contained the GII.P4 ORF1. However, the most recent pandemic GII.4 strain, Sydney 2012, circulated most commonly with the GII.Pe ORF1 and less frequently with the GII.P4 ORF1 (van Beek et al., 2013; Wong et al., 2013). Therefore any substitutions in ORF1 that are essential for pandemicity must have either been acquired by the common ancestor of the GII.P4 and GII.Pe genotypes, or have been acquired convergently along the lineages leading to the pandemic GII.P4 and GII.Pe clades. GII.P4 and GII.Pe last shared a common ancestor in August 1972 (95% HPD October 1966-March 1978). We do not find evidence of convergent changes leading to the pandemic clades within the GII.P4 and GII.Pe genotypes (Figure 2.8). Additionally, while several substitutions occur convergently leading to GII.P4, GII.Pe or the pandemic subclades within these genotypes (Figure S2.4), the residues at these sites within the pandemic clades are additionally found in other genotypes (Figure S2.5) and so are unlikely to have driven pandemic spread. Therefore substitutions within ORF1 required for pandemicity were most likely acquired by the early 1970s and so are unlikely to have driven the increase in prevalence of the GII.4 lineage in the mid-1990s.

2.4.4 An increase in substitution rate leading to the GII.P4 lineage

It has previously been suggested that the high mutation rate associated with the GII.P4 RdRp has contributed to the high prevalence of the GII.4 capsid (Bull et al., 2010). If the underlying mutation rate is different for different RdRps, this would likely leave a pattern in the RdRp phylogenetic tree, with those viruses with an increased mutation rate accumulating greater nucleotide change through time. In a RdRp phylogeny containing sequences from all norovirus genogroups, a group of GII sequences have accumulated more change relative to their sampling date compared with the rest of the GII clade (Figure 2.9). We confirmed this using a phylogenetic tree reconstructed on the GII clade with a single outgroup species (Figure 2.9). The sequences that have accumulated more change form a monophyletic clade within the RdRp phylogenetic tree that includes the

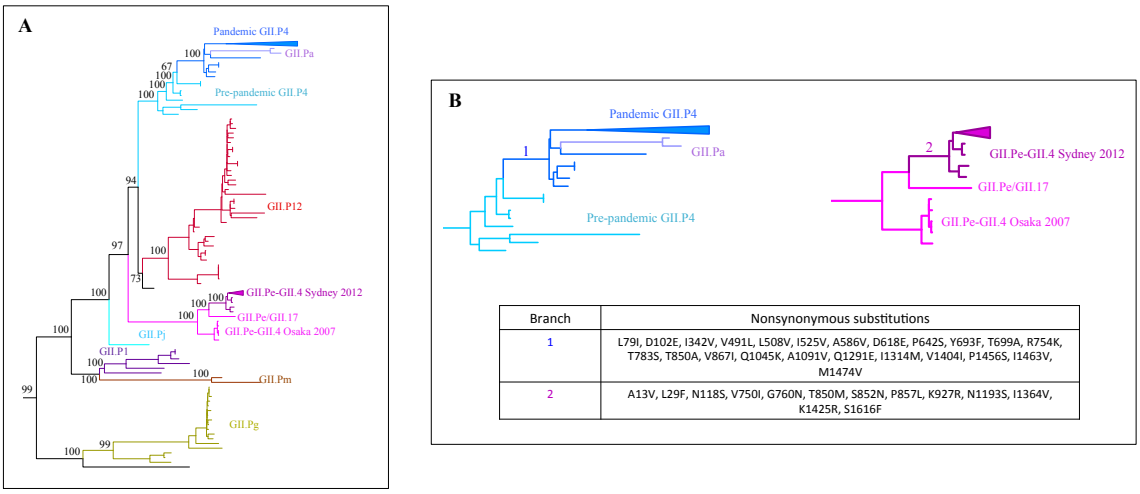


Figure 2.8: Substitutions leading to the pandemic-associated ORF1 regions. (**A**) A nucleotide maximum likelihood phylogenetic tree was reconstructed on all available ORF1 sequences from genogroups GII and GIV. Nonsynonymous substitutions were inferred via ancestral reconstruction. Only a subclade from the tree is shown for clarity and branches within this subclade are coloured by genotype or strain, with the colour matching that of the genotype or strain label. Part of the pandemic GII.P4 clade has been collapsed for clarity. Bootstrap supports are shown on trunk nodes. The GII.Pe sequences are divided into those found with either the GII.4 Osaka 2007 capsid or the GII.17 capsid (both non-pandemic) and those found with the GII.4 Sydney 2012 capsid (pandemic). (**B**) The lineages close to the pandemic GII.P4 clade and the pandemic GII.Pe-GII.4 Sydney 2012 clade are shown. The nonsynonymous substitutions that occurred along the branch leading to the pandemic GII.P4 clade (branch 1) and along the branch leading to the pandemic GII.Pe-GII.4 Sydney 2012 clade (branch 2) are shown.

genotypes GII.P1, GII.P4, GII.P12, GII.Pa, GII.Pc, GII.Pe, GII.Pg, GII.Pj and GII.Pm (Figures 2.9, 2.10). This clade therefore contains each of the RdRp genotypes that are associated with the GII.4 capsid (GII.P1, GII.P4, GII.P12 and GII.Pe) as well as several 'orphan' genotypes, so called because they are not the dominant RdRp found with any of the capsid genotypes (Bull and White, 2011; Kroneman et al., 2013). We therefore named this monophyletic clade the GII.P4 lineage. The RdRp tree suggests that the GII.Pm genotype evolved from GII.P1 viruses that persisted through to the 1990s and 2000s, while the GII.Pa genotype evolved from GII.P4 US95/96 viruses that persisted through to 2010 (Figure 2.10). The GII.P4 lineage may have accumulated more change due to either a higher substitution rate within the clade or a single long branch leading to the common ancestor of the clade. We infer that the substitution rate within the GII.P4 lineage is significantly higher than that in the rest of the GII clade ($p < 0.001$ Kolmogorov Smirnov test), with a general elevation in the substitution rate along branches within the GII.P4

lineage (Figure 2.11). We estimate the mean substitution rate within the GII.P4 lineage to be 6.30×10^{-3} compared with 4.74×10^{-3} in the rest of the GII clade. The substitution rate is determined by a combination of mutation rate, replication dynamics and transmission dynamics and can also be influenced by selection, with positive selection increasing the substitution rate and purifying selection reducing the substitution rate. The GII.P4 lineage exhibits a higher substitution rate at the third codon position compared with the rest of the GII clade (8.24×10^{-3} in the GII.P4 lineage versus 5.68×10^{-3} in the rest of the GII clade, $p < 0.001$ Kolmogorov Smirnov test), suggesting that at least part of the difference in substitution rate is due to a neutral process, such as an increase in mutation rate, faster replication rate or increased transmission rate.

The common ancestor of the GII.P4 lineage occurred in 1906 (95% HPD 1872-1937, Figure 2.10) and it is therefore likely that the increase in substitution rate had been acquired by this date. We would expect that sites in the RdRp that may be responsible for the increase in substitution rate would change leading to the common ancestor of the GII.P4 lineage. We infer that four nonsynonymous substitutions occurred along the branch in the phylogenetic tree leading to the GII.P4 lineage: Y105F, S189A, F426S and S463T (Figure 2.12). Additionally, we identify four sites that are evolving under different selective constraints within the GII.P4 lineage compared with the rest of the GII clade: 105, 266, 330 and 463. Site 105 is completely conserved as phenylalanine in the GII.P4 lineage and completely conserved as tyrosine in the rest of the GII clade (Figure 2.12), suggesting there is a strong selective constraint to retain the respective residue within each clade. Site 189 is also highly conserved as alanine within the GII.P4 lineage, with the rest of the GII clade having either glycine or serine at this site (Figure 2.12). Sites 426 and 463 are variable within the GII.P4 lineage. Site 426 is mostly serine in the GII.P4 lineage, with serine also being found in the GII.P3, GII.P13 and GII.P17 genotypes. Site 463 varies between alanine, serine and threonine within the GII.P4 lineage, with the rest of the GII clade having either alanine, asparagine or serine at this site. Sites 266 and 330 are variable within the GII.P4 lineage and exhibit residues within the GII.P4 lineage that are also present in the rest of the GII clade (Figure 2.12). Therefore the substitutions at sites 266, 330, 426 and 463 are unlikely to have been responsible for the increase in substitution rate leading to the GII.P4 lineage. None of the sites that either change leading to the common ancestor of the GII.P4 lineage or are under different selective constraints in the GII.P4

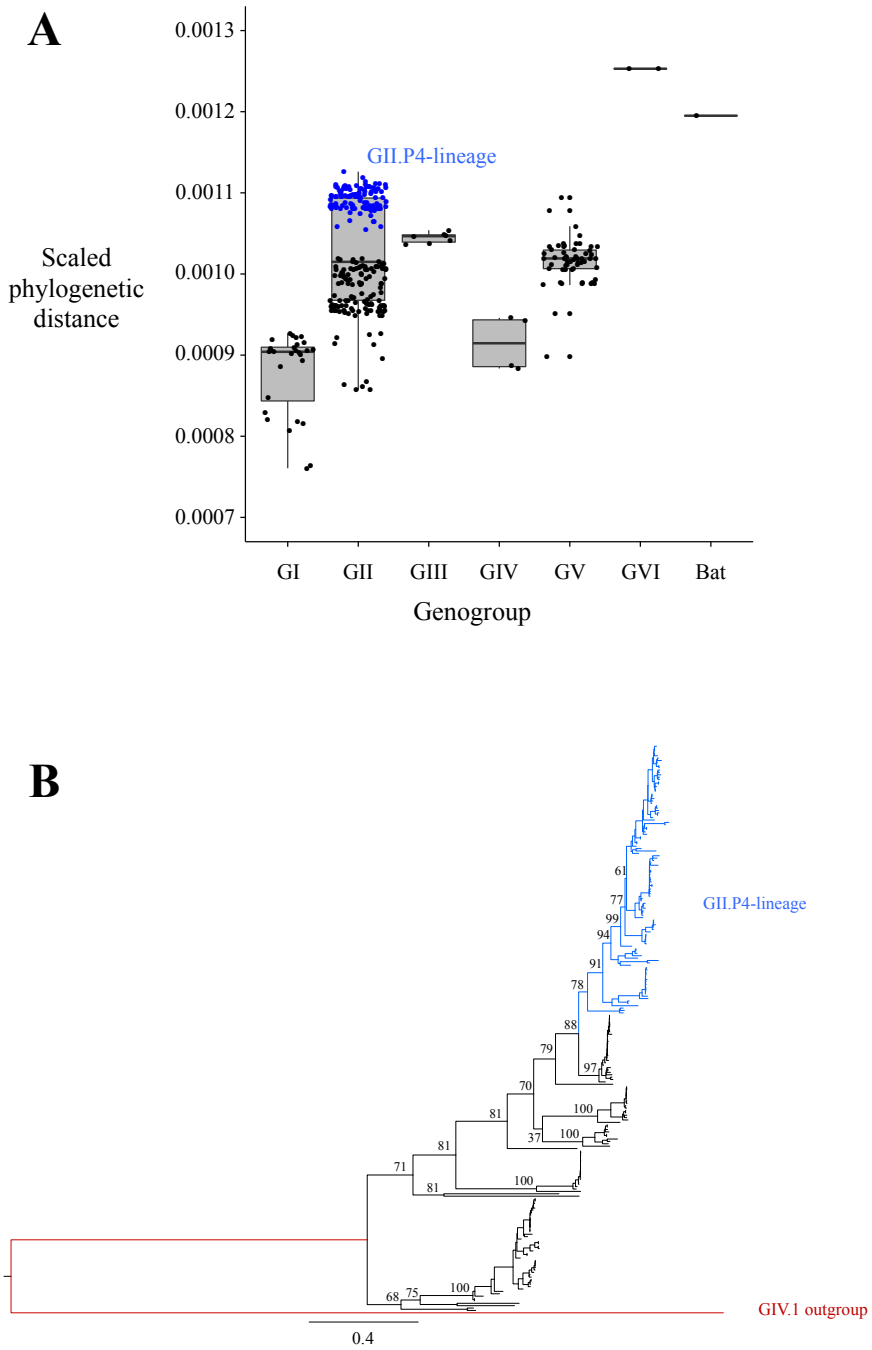


Figure 2.9: The GII.P4 lineage exhibits an excess of change compared with the other GII genotypes. (A) The scaled phylogenetic distance was calculated for each sequence in an all genogroups RdRp phylogenetic tree by dividing the root-to-tip distance by the collection date of each tip to give a measure of accumulation of nucleotide change per unit time. The sequences within the GII.P4 lineage are labelled in blue. **(B)** A nucleotide maximum likelihood phylogenetic tree of the GII RdRp with a GIV.1 outgroup. The GII.P4 lineage is indicated in blue. Bootstrap supports are shown at trunk nodes. The scale bar represents the expected nucleotide substitutions/site. In panels A and B, a dataset where the GII.P4 and GII.Pe-GII.4 Sydney 2012 sequences have been randomly subsampled is shown for clarity. The same results are obtained using the complete dataset.

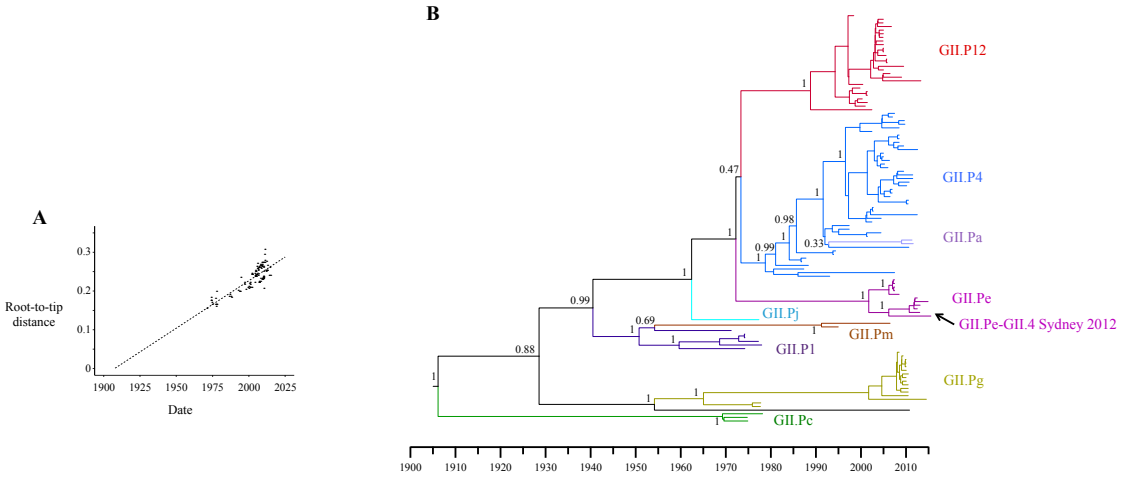


Figure 2.10: Temporal evolutionary history of the GII.P4 lineage. **(A)** Correlation between root-to-tip distance and collection date in a nucleotide maximum likelihood tree of the GIII.P4 lineage. The correlation enabled reconstruction of the temporal evolutionary history of the lineage. **(B)** MCC tree of the GII.P4 lineage reconstructed using BEAST. Branches are coloured by RdRp genotype with the colour matching that of the corresponding genotype label. Posterior supports are shown on trunk nodes.

lineage relative to the rest of the GII clade are close to the active site of the RdRp (Figure 2.13). Sites 105 and 189 are located close together in the fingers domain of the RdRp (Figure 2.13).

2.4.5 VP2 substitutions leading to the pandemic GII.4 lineage

As in the capsid phylogenetic tree, the VP2 regions found with each of the GII.4 strains associated with pandemics and large epidemics since the mid-1990s evolve from a single common ancestor that diverged from the Camberwell 1994-like sequences (Figure 2.14). Therefore, as with the capsid, the GII.4 sequences can be divided into the pandemic GII.4 clade and the pre-pandemic GII.4 lineages. The common ancestor of the pandemic GII.4 clade in the VP2 region occurred in July 1992 (95% HPD February 1991-March 1994, Figure 2.14), the same month as the capsid common ancestor (Figure 2.2). The pandemic GII.4 clade diverged from the pre-pandemic GII.4 viruses in October 1988 (95% HPD January 1988-July 1990). Ten nonsynonymous substitutions occurred along the branch leading to the common ancestor of the pandemic GII.4 clade: Q72K, L78M, T93E, V148A, P155S, P156S, A162T, G187S, N199S and N230S (Figure 2.14).

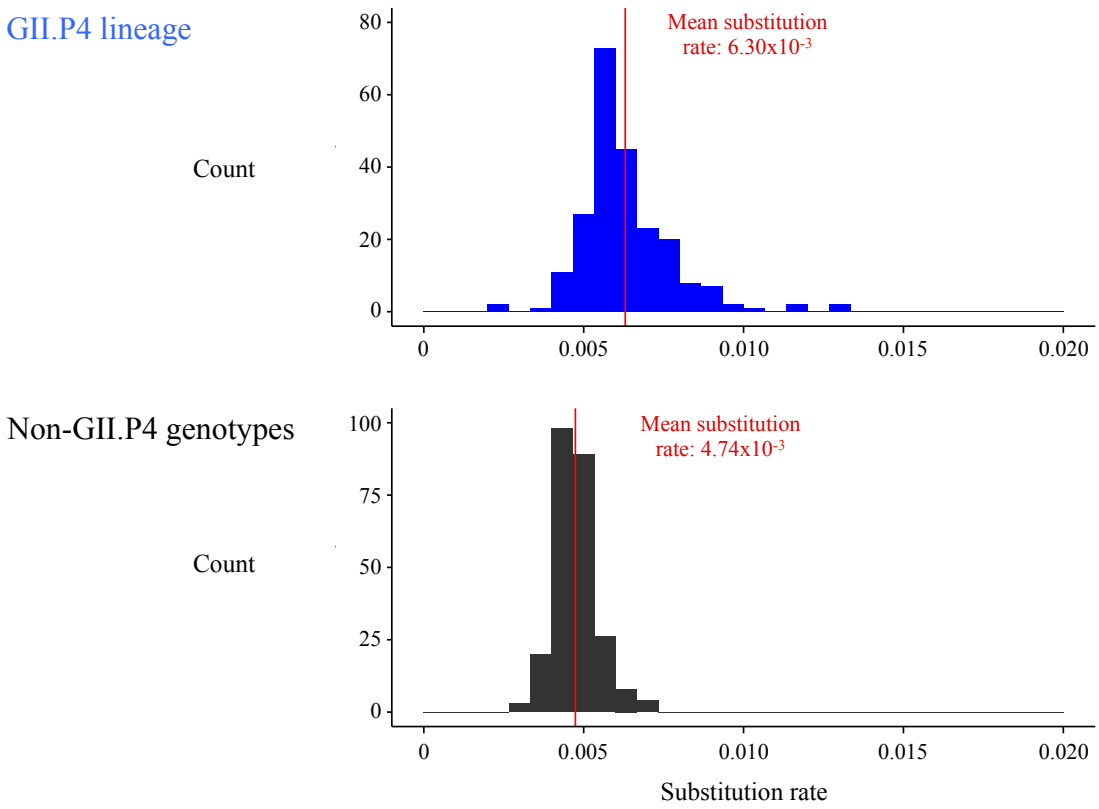


Figure 2.11: The GII.P4 lineage has an elevated substitution rate compared with the rest of the GII clade. A MCC tree was reconstructed for the GII.P4 lineage and the rest of the GII clade using BEAST under a relaxed lognormal clock model. The distribution of branch substitution rates is plotted for the respective MCC trees. The mean substitution rate estimated for each dataset from the complete posterior distribution of trees is indicated with a red vertical line.

However, sites 155 and 199 undergo a back substitution along the branch leading to the common ancestor of the pandemic GII.4 clade, with the dominant residue within the pandemic GII.4 clade matching that in the early pre-pandemic GII.4 sequences (Figure 2.15). It is therefore unlikely that these sites contributed to the increase in frequency of the GII.4 genotype. Additionally, site 156 exhibits the same residue in the pre-pandemic GII.4 sequences as in the Asia 2003 epidemic strain and a large clade within the Den Haag 2006 pandemic strain, while site 162 exhibits the same residue in the pre-pandemic GII.4 sequences as in the Hunter 2004 and Den Haag 2006 pandemic strains (Figure 2.15). Site 230 exhibits the same residue in the pandemic GII.4 clade as in the Bristol 1993 pre-pandemic clade (Figure 2.15). It is therefore unlikely that any of these sites were responsible for the increase in frequency of the GII.4 genotype in the mid-1990s. Sites 72, 78 and 93 are largely conserved as K, M and E, respectively within the pandemic GII.4

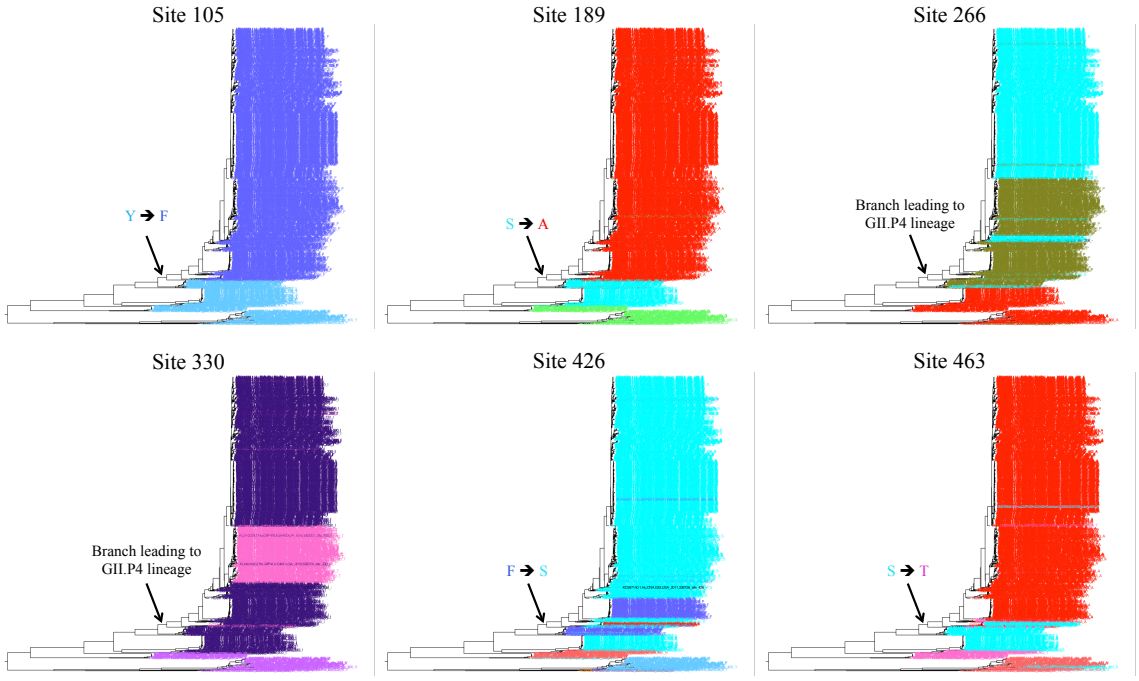


Figure 2.12: Sites changing leading to and under different selective constraints within the GII.P4 lineage. Coloured trees for sites that change leading to the GII.P4 lineage (105, 189, 426 and 463) and sites that are under different selective constraints within the GII.P4 lineage compared with the rest of the GII clade (105, 266, 330 and 463). The branch leading to the common ancestor of the GII.P4 lineage is indicated in each tree and the substitution along that branch is indicated for the four sites that change along that branch.

clade, with changes away from the dominant residue in the pandemic GII.4 clade typically only occurring along tip branches within the tree, or leading to small clades containing viruses that were sampled over a short period of time (Figure 2.16). There is therefore no evidence that viruses within the pandemic GII.4 clade with changes away from the dominant residue at sites 72, 78 and 93 can persist. This suggests that viruses within the pandemic GII.4 clade have a constraint to retain the dominant residue at these positions. Site 148 is highly variable within the pandemic clade, with alanine, aspartic acid and threonine at this site, compared with the pre-pandemic viruses in which valine is conserved. Site 187 also varies between asparagine, arginine and serine within the pandemic GII.4 clade, each of which differs from the glycine that is dominant in the pre-pandemic GII.4 viruses (Figure 2.16).

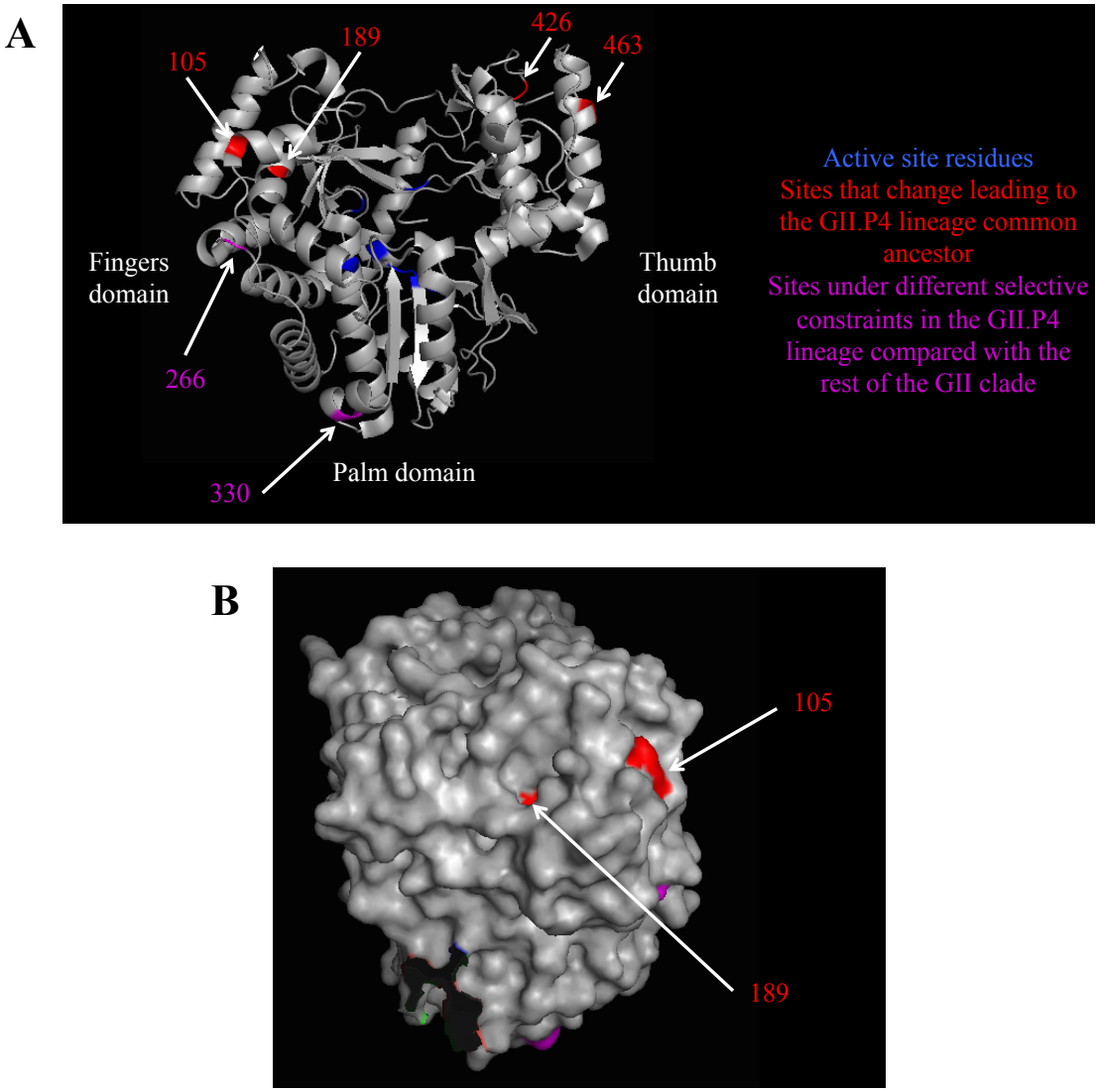


Figure 2.13: Location of sites 105 and 189 in the RdRp structure. **(A)** The RdRp active site residues (blue), sites that change leading to the common ancestor of the GII.P4 lineage (red) and sites under different selective constraints in the GII.P4 lineage compared with the rest of the GII clade (magenta) are shown on the GII.P4 RdRp structure 1SH0. **(B)** Side view of the fingers domain of the RdRp, showing sites 105 and 189 in close apposition.

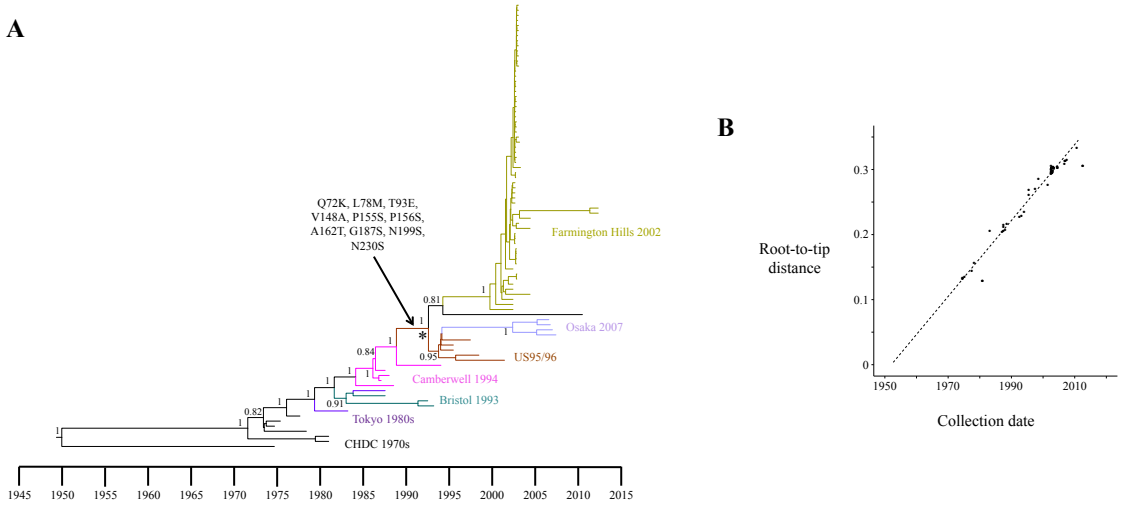


Figure 2.14: Temporal history of the early GII.4 VP2 lineage. (**A**) MCC tree reconstructed using BEAST of the early GII.4 lineage VP2, consisting of the CHDC 1970s, Tokyo 1980s, Bristol 1993, Camberwell 1994-like, US95/96 and Osaka 2007 strains, with Farmington Hills 2002 included as an early strain within the pandemic GII.4 clade. Branches are coloured by strain, with the colour matching that of the corresponding strain label. As there is currently no genotyping system established for VP2, one of the sequences with a Tokyo 1980s capsid clusters within the Bristol 1993 clade within VP2, as shown by the purple branch within the Bristol 1993 clade. The starred node is the common ancestor of the pandemic GII.4 clade. Posterior supports are shown at trunk nodes. The nonsynonymous substitutions that occurred leading to the common ancestor of the pandemic GII.4 clade are labelled. (**B**) The correlation between root-to-tip distance and collection date calculated using a maximum likelihood tree reconstructed on the same sequences as in panel A. The correlation between these values enabled the reconstruction of the temporal evolutionary history of the lineage.

2.4.6 The GII.4 capsid exhibits an accumulation of amino acid change through time

We next investigated whether the evolutionary process within the GII.4 genotype is different to that within other GII genotypes. We reconstructed nucleotide phylogenetic trees of the GII.1, GII.2, GII.3, GII.5, GII.6, GII.7, GII.12, GII.13, GII.14, GII.17 and GII.21 capsid genotypes, in addition to the GII.4 phylogenetic tree used in our previous analyses. Most of the capsid genotypes exhibit strong temporal signal at the nucleotide level, with a correlation between nucleotide root-to-tip distance and collection date (Figure 2.17). The GII.14 capsid exhibits only weak temporal signal, indicating the lack of a steady accumulation of change through time in this genotype (Figure 2.17). The GII.17 capsid exhibits little temporal signal, consistent with previous suggestions that different

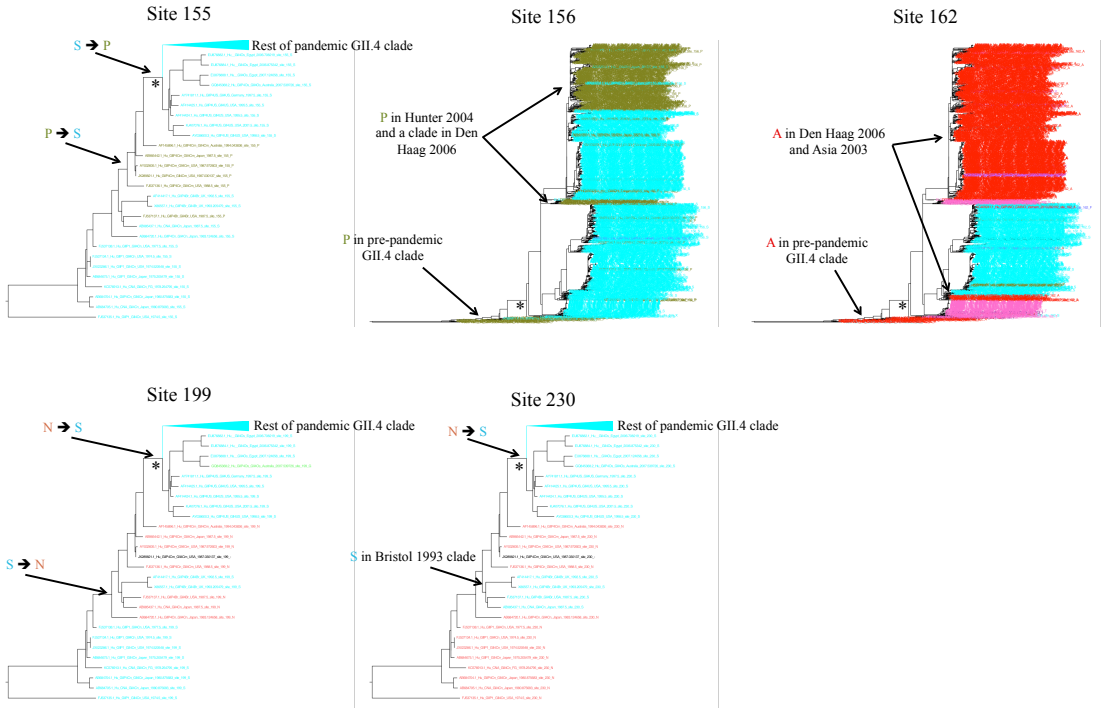


Figure 2.15: VP2 sites that change leading to the pandemic GII.4 clade but were unlikely to have driven an increase in prevalence. Coloured trees of sites 155, 156, 162, 199 and 230. The starred node in each tree is the common ancestor of the pandemic GII.4 clade. The pandemic GII.4 clade after the divergence of the US95/96 and Osaka 2007 strains has been collapsed in the trees for sites 155, 199 and 230 for clarity. Sites 155 and 199 exhibit back substitutions leading to the common ancestor of the pandemic GII.4 clade, with the dominant resiude in the pandemic GII.4 clade also being dominant in the early GII.4 sequences. Sites 156 and 162 undergo back substitutions within pandemic GII.4 strains to the dominant residue in the pre-pandemic GII.4 sequences. The dominant residue in the pandemic GII.4 clade at site 230 was also present in the pre-pandemic Bristol 1993 clade.

GII.17 lineages may evolve at different rates (Lu et al., 2016). The GII.6 genotype also does not exhibit temporal signal when all GII.6 sequences are included (Figure 2.17). However, the GII.6 genotype clusters into three well diverged subclades, with each of these clades being separated by a long branch within the phylogenetic tree (Figure S2.6) (Vinje, 2015). Each of these three subclades does exhibit temporal signal (Figure S2.6), suggesting the lack of temporal signal within the GII.6 genotype as a whole is due to either differences in substitution rate leading to the clade common ancestors, differences in substitution rate within each of the GII.6 clades, or both.

The non-GII.4 genotypes that exhibit a strong temporal signal at the nucleotide level exhibit only a weak temporal signal, or no temporal signal, at the amino acid level and do not exhibit an accumulation of amino acid change through time (Figure 2.18). The

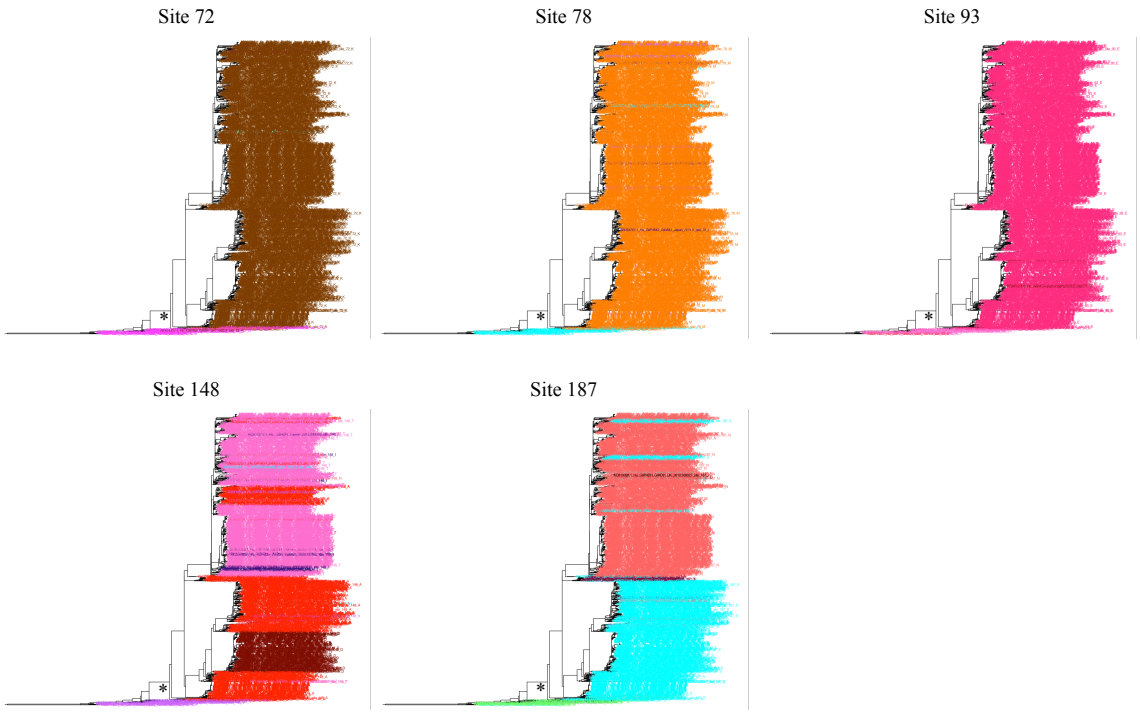


Figure 2.16: VP2 sites that change leading to the pandemic GII.4 clade and exhibit residue differences between the pandemic GII.4 clade and pre-pandemic GII.4 sequences. Coloured trees for sites 72, 78, 93, 148 and 187. The starred node in each tree is the common ancestor of the pandemic GII.4 clade. Sites 72, 78 and 93 are conserved within the pandemic GII.4 clade. Sites 148 and 187 are variable within the pandemic GII.4 clade, with each common residue within the pandemic GII.4 clade being different to that in the pre-pandemic GII.4 sequences.

GII.4 capsid, however, exhibits a strong temporal signal at the amino acid level. This accumulation of amino acid change is evident in both the pandemic GII.4 clade and the pre-pandemic GII.4 lineages (Figure 2.18), although the magnitude of the accumulation appears greater within the pandemic GII.4 clade (Figure S2.7).

2.5 Discussion

The GII.4 capsid genotype has dominated norovirus outbreaks since the mid-1990s and during this time has caused six pandemics and additional epidemics (Siebenga et al., 2009; van Beek et al., 2013). Our results support previous suggestions that the pandemic caused by US95/96 was the first GII.4 pandemic and coincided with a large increase in prevalence of this genotype (Figure 2.2) (Donaldson et al., 2008; Siebenga et al., 2010), although reconstruction of the evolutionary dynamics of the early GII.4 lineage suggests that this first pandemic may have begun as early as 1994 (Figure 2.2 panels C, D). The

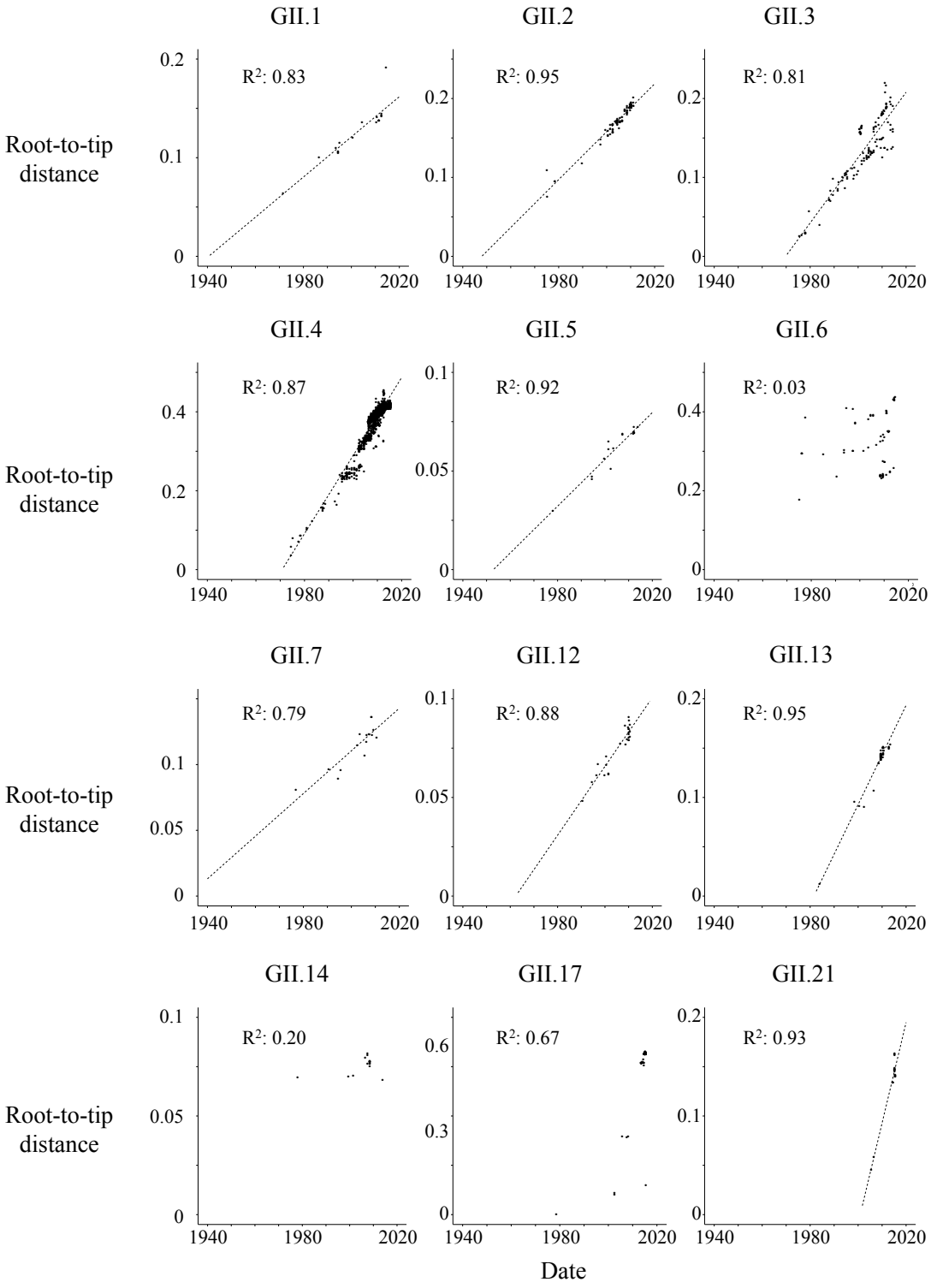


Figure 2.17: Accumulation of nucleotide change within GII capsid genotypes. A nucleotide maximum likelihood phylogenetic tree was reconstructed for each genotype. Here, the nucleotide root-to-tip distance in the phylogenetic tree is plotted against the collection date for each sequence. The root location was identified as the position that minimised the heuristic residual mean squared of the root-to-tip distance versus collection date, as calculated with TempEst v1.5 (Rambaut et al., 2016). The dashed line is a linear regression between root-to-tip distance and collection date and is shown for genotypes that exhibit a correlation between these values. The R^2 correlation between root-to-tip distance and collection date is shown for each genotype.

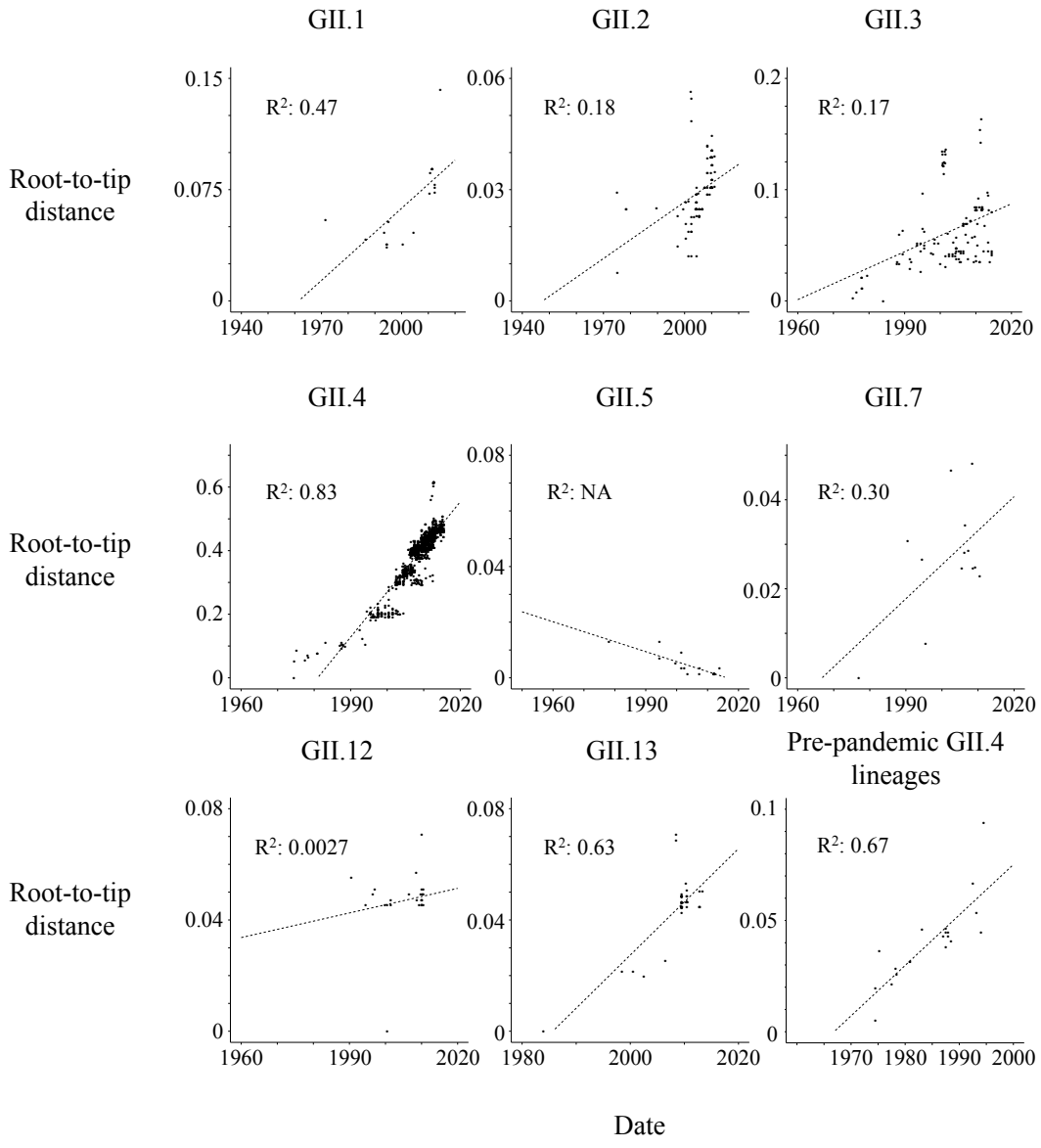


Figure 2.18: Accumulation of amino acid change within GII capsid genotypes. The amino acid root-to-tip distance versus collection date is plotted for each sequence within each genotype that exhibited a correlation between nucleotide root-to-tip distance and collection date. The dashed line is a linear regression between root-to-tip distance and sampling date. The R^2 correlation between amino acid root-to-tip distance and collection date is shown for each genotype. As the GII.5 genotype exhibits a negative correlation between amino acid root-to-tip distance and collection date, the R^2 correlation was not calculated for this genotype.

timing of the first pandemic has the caveat that there was an increase in norovirus sampling in the mid-1990s that may have resulted in an artificial increase in the estimated relative genetic diversity at this time. It is unlikely that a change in the host population could have selectively driven the emergence of only a single norovirus genotype. Instead, the transition of GII.4 from one of a number of low prevalence genotypes to the dominant genotype is most likely to have been driven by one or more changes in the viral genome. Importantly, the continued high prevalence of GII.4 through to the present day suggests that the change(s) important for the increase in frequency have been retained, at least phenotypically if not genotypically. The structure of the GII.4 capsid and VP2 phylogenetic trees demonstrates that each of the strains associated with pandemics or large epidemics since the mid-1990s cluster downstream of a single node (Figures 2.1, 2.14). Therefore GII.4 transitioned from being one of a number of low frequency genotypes to being the dominant genotype along a single branch within the respective phylogenetic trees. Conversely, the ORF1 regions found with pandemic strains last shared a common ancestor in the 1970s (Figure 2.10). Therefore there was not a single branch in the ORF1 phylogenetic tree where the transition from low frequency to high frequency occurred and we find no evidence of convergent changes occurring along branches leading to high frequency ORF1 clades (Figures 2.8, S2.4, S2.5). This has the caveat that changes at different sites within ORF1 proteins may result in phenotypically similar outcomes; this is hard to assess with the currently limited knowledge of the functions of ORF1 proteins and functionally important sites. However, our results argue that genetic changes within the capsid and/or VP2 drove the increase in frequency of GII.4. The capsid and VP2 substitutions that occurred leading to the pandemic GII.4 clade were most likely acquired by July 1992 (Figures 2.2, 2.14) and therefore GII.4 likely circulated with these changes for at least two years prior to increasing in frequency.

Our results suggest two potential mechanisms by which the GII.4 capsid could have increased in prevalence in the mid-1990s: an increase in capsid stability and an increased range of HBGA-binding. Sites 333 and 459 were the only two sites identified as evolving under different selective constraints in the pandemic GII.4 clade and the pre-pandemic GII.4 lineages, although the small number of sequences in the pre-pandemic GII.4 clade raises the possibility of false negatives. We infer that the leucine to glutamine substitution at site 459 may have increased the interaction network within region IV of the capsid

dimer interface (Figure 2.5, S2.2); future crystallographic studies to examine the structure in this region with leucine at site 459 would be useful to confirm this. The potential influence of the substitution at site 333 is less clear, although the methionine and valine at this site in the pandemic GII.4 clade (Figure 2.4) have greater beta sheet propensity than the leucine found in the pre-pandemic clade (Koehl and Levitt, 1999) and may therefore alter local protein structure or flexibility. Alternatively, these residues may increase dimer stability through specific inter-monomer interactions. Clearly, an increase in inter-monomer interactions is likely to increase the stability of the viral particle. A more stable capsid would be able to persist for longer on surfaces, in the environment and within an infected individual, resulting in significantly increased transmission. To date, studies directly comparing the environmental stability of different norovirus genotypes are lacking. However, there is evidence that the Ebola virus responsible for the large 2013-2016 epidemic was more stable in blood compared with a 1976 Ebola virus that caused only a small outbreak and there is therefore precedent for more stable variants causing larger outbreaks (Schuit et al., 2016).

While the majority of GII capsid genotypes exhibit an accumulation of nucleotide change through time (Figure 2.17), GII.4 was the only genotype to exhibit an accumulation of amino acid change through time (Figure 2.18). These results are consistent with a concurrent published study that demonstrated that non-GII.4 genotypes remain 'static' at the amino acid level (Parra et al., 2017). Interestingly, the ability to accomodate change appears to be inherent to the GII.4 capsid, as the pre-pandemic GII.4 lineages also exhibit an accumulation of amino acid change through time (Figure 2.18). However, the magnitude of this accumulation appears far greater in the pandemic GII.4 clade (Figure S2.7), indicating an elevated rate of change following pandemic emergence. It is not clear whether this increase is a cause or consequence of pandemic spread. Future analyses estimating substitutions rates and dN/dS ratios for the pandemic GII.4 clade and pre-pandemic GII.4 lineages would be useful to confirm our findings. It has previously been demonstrated that the increased mutational robustness conferred by an increase in protein stability increases the capacity of a protein to evolve (Bloom et al., 2006). Capsid sites 93, 172, 176, 285, 497 and 505 changed leading to the pandemic GII.4 clade and are conserved within this clade (Figure 2.7). While these sites are not currently associated with a particular function, their conservation within the pandemic GII.4 clade suggests

a selective constraint. We therefore hypothesise that one or more of these substitutions, potentially coupled with the substitutions at site 333 and/or 459, may have increased capsid stability and thereby increased evolvability, enabling accomodation of the amino acid change necessary to persist at high frequency within the population. Future *in silico* studies examining the adaptability of the capsid protein with and without these substitutions would test this hypothesis. Additionally, future studies comparing the environmental stability of pandemic and pre-pandemic GII.4 capsids would resolve whether there was indeed an increase in stability leading to the pandemic GII.4 clade.

Previous studies have demonstrated that the US95/96 GII.4 strain binds to a greater range of HBGAs than other genotypes, providing a mechanism by which this strain can infect a larger proportion (roughly 80%) of the population (Lindesmith et al., 2008; Donaldson et al., 2008). Site 395 is the only site close to the HBGA-binding region that changed leading to the common ancestor of the pandemic GII.4 clade (Figure 2.2) and it is therefore likely that any alterations in HBGA binding between the Camberwell 1994-like and US95/96 strains would be driven by the substitution at this site. Movement of the 391-395 loop is required for binding to HBGA type B, Le^B and Le^Y (Singh et al., 2015) and substitutions at site 395 can alter binding to HBGA type A (Lindesmith et al., 2008). Therefore substitutions at site 395 have the potential to alter HBGA binding by both altering movement of the 391-395 loop and by directly modulating interactions with HBGAs. Each of the amino acid residues found at site 395 in the pandemic GII.4 clade (alanine, asparagine and threonine) is smaller in size than the histidine found in the pre-pandemic GII.4 lineages (Figure 2.4). Interestingly, site 395 is also part of blockade epitope D (Lindesmith et al., 2012a) and substitutions at this site within the pandemic GII.4 clade that enable evasion of antibody responses have retained the small residue size at this site. We hypothesise that a smaller residue at site 395 results in a more flexible loop and therefore more efficient binding to a greater range of HBGAs, enabling infection of a larger proportion of the population. While there is not a clear difference in the structure of the 391-395 loop in homology models of the pandemic GII.4 clade common ancestor and the immediately upstream ancestor, under our hypothesis it is movement of this loop that is important and such movement is not included in the homology models. Therefore future structural, molecular dynamics and HBGA-binding studies would determine whether the size of the residue at site 395 alters movement of this loop and HBGA-binding. However,

the role of wider HBGA-binding in the high prevalence of GII.4 is called into question by the low efficiency of binding of the Hunter 2004 pandemic strain to synthetic HBGAs in experimental conditions (Lindesmith et al., 2008, 2012a). If wider HBGA-binding is the key mechanism enabling the high prevalence of GII.4, it would be expected that this be maintained in all pandemic GII.4 strains. It is, however, possible that Hunter 2004 may bind to alternative cell surface receptors.

While sites 294, 333, 340 and 395 are located within blockade epitope A, putative epitope B, putative epitope C and blockade epitope D, respectively, envoking evasion of immunity as a mechanism by which the GII.4 capsid increased in prevalence would require significant population immunity built against pre-pandemic GII.4 viruses from which the pandemic GII.4 lineage escaped. Given the low prevalence of GII.4 viruses prior to the mid-1990s it is very unlikely that such immunity existed.

The influence of the substitutions at sites 72, 78, 93, 148 and 187 in VP2 is difficult to predict due to the lack of knowledge of the functional importance of different VP2 regions. While site 148 is located within the putative capsid-interaction domain (Glass et al., 2003), variation at this site does not correlate with variation within the capsid shell domain and is therefore unlikely to greatly alter capsid-VP2 interaction.

The increase in prevalence of the GII.4 capsid is unlikely to have been driven by an increase in mutation rate (Bull et al., 2010) as, while we do find evidence of an elevated substitution rate in the GII.P4 lineage, this higher rate was likely acquired by 1906 (Figure 2.10) and was therefore not newly acquired close to the time of the increase in GII.4 prevalence. Importantly, the accumulation of nucleotide change is relatively constant throughout the GII.P4 lineage (Figure 2.10 panel A) and this lineage exhibits a generally elevated substitution rate compared with the rest of the GII clade (Figure 2.11). The higher substitution rate at the third codon position within the GII.P4 lineage suggests at least part of the increased substitution rate is due to a neutral process, such as an increase in mutation rate, replication rate or transmission rate. Several of the genotypes within the GII.P4 lineage are found with the GII.4 capsid which exhibits a rapid transmission rate, as evidenced by its high frequency. However, the presence of the rarely detected ‘orphan’ ORF1 genotypes within this lineage (Figure 2.10) suggests that increased transmission is not the factor driving the higher substitution rate. Additionally, the replication rate of the GII.P4 RdRp is not higher than that of the GII.Pb or GII.P7 RdRps (genotypes that

are not within the GII.P4 lineage) (Bull et al., 2010). We therefore hypothesise that the increased substitution rate is due to an increase in mutation rate. Future studies investigating the effect of the substitutions at RdRp sites 105 and 189 on the mutation rate and replication rate using previously developed assays (Bull et al., 2010) would confirm this. It was recently demonstrated that a more diverse viral quasispecies results in more efficient transmission of murine norovirus (Arias et al., 2016) and therefore an increased mutation rate generating a more diverse viral population may enable more efficient transmission between humans. Each of the ORF1 genotypes found with the pandemic (and pre-pandemic) GII.4 strains is within the GII.P4 lineage (Figure 2.10) and a higher mutation rate may therefore contribute to the rapid transmission of GII.4, although it cannot be completely responsible for this rapid transmission.

We hypothesise that the GII.4 genotype was pre-adapted to become pandemic due to its capsid structure inherently being able to tolerate high levels of amino acid variation and through the associated RdRps from the GII.P4 lineage enabling rapid mutation to enable efficient transmission and evasion of host immunity. We propose that this combined with an increase in capsid stability and/or increase in HBGA binding in the mid-1990s resulted in a highly stable and highly transmissible virus that was capable of infecting a large proportion of the human population and tolerating the variability that has been required to persist despite significant population immunity. The combination of these factors has enabled the GII.4 genotype to dominate gastroenteritis cases and outbreaks worldwide for the past 20 years.

Chapter 3

Identifying the sources and drivers of norovirus pandemics

3.1 Abstract

The norovirus genotype GII.4 is a leading cause of human gastroenteritis and has caused six pandemics since the mid-1990s. Hitherto it has been suggested that pandemics are driven by the acquisition of viral mutations enabling evasion of existing herd immunity. However, by reconstructing the evolutionary dynamics of GII.4 norovirus, we demonstrate that all of the GII.4 strains circulated undetected for years prior to their pandemic spread. We show that each of the pandemic GII.4 strains diverged into multiple lineages over years prior to pandemic emergence, indicating that the genetic changes important for pandemic spread were acquired several years previously. Our results suggest that GII.4 strains circulate in their pandemic form within one or more currently poorly sampled reservoir populations for years prior to pandemic spread. We hypothesise that new pandemics are triggered when growing herd immunity against the preceding pandemic strain opens a niche into which a pre-adapted GII.4 strain can emerge. These results have implications for surveillance strategies to detect potential future pandemic strains and for vaccine design.

3.2 Introduction

Despite the large number of norovirus genotypes known to infect humans, all pan-

demics and up to 80% of annual outbreaks since 1995 have been caused by a single capsid genotype, GII.4 (Kroneman et al., 2008; Siebenga et al., 2009). To date, there have been six GII.4 pandemics, occurring in the Northern hemisphere winters of 1995-1996 (US95/96 strain), 2002-2003 (Farmington Hills 2002), 2004-2005 (Hunter 2004), 2006-2007 (Den Haag 2006), 2009-2010 (New Orleans 2009) and 2012-2013 (Sydney 2012) (Siebenga et al., 2009; van Beek et al., 2013). Each pandemic has been caused by a phylogenetically distinct GII.4 strain and the replacement of the previous pandemic strain occurs rapidly, with the new strain dominating outbreaks worldwide within several months of its emergence (Siebenga et al., 2007, 2009; White, 2014). Here, we define a pandemic strain as the dominant strain in worldwide outbreaks. There are therefore multiple pandemic strains within the 'pandemic GII.4 clade' defined in the previous chapter and replacement of the pandemic strain has occurred periodically since the mid-1990s. Several epidemic GII.4 strains have also been identified and have caused outbreaks either within more restricted geographical regions, or have caused widespread but only sporadic outbreaks, including Asia 2003, Yerseke 2006, Apeldoorn 2007 and Osaka 2007 (Siebenga et al., 2009; Eden et al., 2013).

It has previously been suggested that, analogous to influenza A H3N2, the pandemic GII.4 strains generally evolve from one of the previous pandemic GII.4 strains (Siebenga et al., 2007, 2009; van Beek et al., 2013; White, 2014), a process driven by mutations in the antigenic regions of the capsid P2 domain that enable evasion of population herd immunity (Lindesmith et al., 2008; Cannon et al., 2009; Lindesmith et al., 2012a; Zakikhany et al., 2012; Lindesmith et al., 2013; Debbink et al., 2013). The emergence of new pandemic strains being driven by evasion of herd immunity is supported by the observation that newly emerging pandemic strains are capable of evading mouse monoclonal antibodies, human monoclonal antibodies, mouse polyclonal sera and human polyclonal sera raised against the preceding pandemic strain (Lindesmith et al., 2008, 2011, 2012a, 2013; Debbink et al., 2013). However, recent phylogenetic analysis has suggested that the two most recent pandemic strains, New Orleans 2009 and Sydney 2012, diverged at the start of the Apeldoorn 2007 epidemic and therefore neither evolved from a previous pandemic strain (Eden et al., 2014). Further retrospective analyses have identified sporadic community cases and/or outbreaks caused by pandemic norovirus strains several years prior to the pandemic emergence of those strains (Siebenga et al., 2009; Sdiri-Loulizi et al., 2009;

Eden et al., 2013; Allen et al., 2016). This has led to suggestions that variants of pandemic strains may persist at low levels until acquiring changes in the capsid P2 domain that drive pandemic spread (Eden et al., 2014; White, 2014; Allen et al., 2016). While most work on the drivers of GII.4 pandemics has focused on the capsid, both New Orleans 2009 and Sydney 2012 acquired new ORF1 regions by recombination, leading to suggestions of a role for recombination in strain emergence (Motomura et al., 2010; Eden et al., 2013).

Here, we further investigate the origin of pandemic GII.4 viruses by reconstructing the evolutionary history of the RdRp, capsid and VP2. We demonstrate that each of the pandemic and epidemic GII.4 strains circulated undetected for years prior to pandemic or epidemic spread. The simultaneous presence of multiple unobserved co-circulating GII.4 lineages, including future pandemic and epidemic strains, suggests that these viruses can persist for long periods of time in unsampled reservoir populations. During this time, the pandemic strains diversify into multiple lineages, with these multiple lineages emerging simultaneously at the onset of the new pandemic. This demonstrates that the genetic changes enabling pandemic spread must be acquired years prior to the start of the pandemic and emergence is therefore not driven by the acquisition of new viral mutations. Instead, we hypothesise that new GII.4 pandemics are driven by changes in host herd immunity that create an immunological niche that allows a pre-adapted strain to emerge.

3.3 Materials and Methods

3.3.1 Analyses using all GII.4 strains

To examine the evolutionary dynamics of the GII.4 genotype, we collected all of the 871 GII.4 norovirus sequences available in GenBank as of 30/10/2015 containing the complete RdRp, capsid and VP2. The strategy of including sequences containing each of these regions enabled collection of comparable datasets for each genomic region. Sequences were removed if their GenBank record was associated with terms that may indicate sequence change after collection from an infected patient (as detailed in chapter 2.3.1) and sequences were retained only if they were associated with a collection date in either their GenBank record or within primary literature. The RdRp and capsid genotype

Strain name	Number of RdRp sequences	Number of capsid and VP2 sequences	Subsampled
GII.4 1970s (GII.P1 RdRp)	6	6	No
Bristol 1993	2	2	No
Camberwell 1994	5	5	No
US95/96	3	3	No
Lanzhou 2001	1	1	No
Farmington Hills 2002	18	18	No
Asia 2003 (GII.P12 RdRp)	15	15	No
Hunter 2004	17	17	No
Yerseke 2006	12	12	No
Den Haag 2006	551	559	Yes, 41 sequences
Osaka 2007 (GII.Pe RdRp)	5	5	No
Apeldoorn 2007	37	31	No
New Orleans 2009	141	134	Yes, 41 sequences
Sydney 2012 (GII.Pe RdRp)	36	41	No
GII.4 could not assign strain	3	3	No

Table 3.1: Summary of the GII.4 strains included in this study and the number of sequences from each strain. The number of sequences from each strain is shown following the removal of potentially recombinant samples. The RdRp genotype is shown in parentheses for each GII.4 strain that is not found with the GII.P4 RdRp. Den Haag 2006 and New Orleans 2009 were randomly subsampled to the number of sequences present in the third most prevalent strain. Sydney 2012 and Osaka 2007 are found with the GII.Pe RdRp, therefore there are 41 GII.Pe samples in the RdRp dataset. Pandemic strains are labelled in bold font.

of each sequence was confirmed using the norovirus genotyping tool (Kroneman et al., 2011), which was also used for initial strain typing (Table 3.1). All capsid sequences in the dataset clustered with a bootstrap score of 100 in a capsid phylogeny including viruses from all GII genotypes indicating that each sequence contains the GII.4 capsid. However, the dataset includes sequences with the GII.P1, GII.P4, GII.P12 and GII.Pe RdRp, due to intergenotype recombination events (Table 3.1) (Eden et al., 2013). The strain names used are those returned by the norovirus genotyping tool (Table 3.1). The RdRp, capsid and VP2 datasets were aligned independently at the amino acid level using MUSCLE (Edgar, 2008).

3.3.2 Recombination analysis

The presence of recombination was screened for using the Single Breakpoint (SBP) method in HyPhy (Kosakovsky Pond et al., 2005; Pond et al., 2006). SBP was run on the alignment for the RdRp, capsid and VP2 separately with the GTR model of nucleotide substitution. No recombination was detected in the RdRp or VP2. Recombination was detected in the capsid ($p < 0.01$ in the Kishino-Hasegawa (KH) test). We identified the most likely breakpoint as the nucleotide position with the greatest gain in AIC when the alignment was split at that position compared with the null model and reconstructed a maximum likelihood tree on either side of this position using RAxML (Stamatakis, 2014) with the GTR model of nucleotide substitution and gamma rate heterogeneity with four gamma classes. We assessed topological robustness using 1000 bootstrap replicates. As SBP identifies putative recombination events to the nearest variable site, the identified breakpoint may not be the true breakpoint position. Sequences that clustered separately on either side of the breakpoint with strong bootstrap support (here defined as 70 or above) were considered putative recombinants and were removed from the dataset and from further analyses. SBP was then run on the alignment again to ensure the signal for recombination at that position had been removed. In total, there were 19 putatively recombinant capsid sequences with the most likely breakpoint at either position 314 or 537 (Table S3.1). These recombination events predominantly occurred between Den Haag 2006 and Apeldoorn 2007, Den Haag 2006 and New Orleans 2009 or Apeldoorn 2007 and New Orleans 2009, strains that cocirculated within the human population (Eden et al., 2014), providing the opportunity for coinfection.

3.3.3 Assessing clock-like signal

Methods employing sampling dates to infer divergence times and evolutionary rates are valid only if there is a temporal evolutionary signal in the dataset, i.e. if there is an accumulation of nucleotide change through time (Rambaut et al., 2016). To assess whether there is a clock-like signal in our datasets, we reconstructed maximum likelihood trees for the RdRp, capsid and VP2 using RAxML as described above. The best fitting root position was inferred using TempEst v1.5 (Rambaut et al., 2016). We calculated the R^2 correlation between root-to-tip distance and sampling date and assessed statistical

significance using a bootstrapping approach whereby the sampling dates were randomly resampled and the R^2 correlation recalculated 1000 times.

3.3.4 Bayesian reconstruction of evolutionary dynamics

We used the Bayesian Markov chain Monte Carlo approach implemented in BEAST version 2.2.1 (Bouckaert et al., 2014) to reconstruct the evolutionary dynamics of the GII.4 genotype. Analyses were run separately on the RdRp, capsid and VP2. As Den Haag 2006 and New Orleans 2009 have a greater number of samples compared with the other strains (Table 3.1), we took 3 random subsamples of 41 sequences (chosen to match the number of sequences from the next most numerous strain in each dataset) from each of these strains and combined these with the sequences from the other strains to form 3 subsampled datasets for each genomic region. These datasets contained 241 sequences for the capsid and VP2 and 242 sequences for the RdRp. Each of the subsampled datasets was analysed separately and results were insensitive to the subsampled dataset (Figure S3.1, Table S3.2). Each sample was labelled with the most accurate date possible; the day of collection if available, the middle of the month of collection if the sampling day was not available or the middle of the year of collection if the sampling month was not available. Each dataset was analysed with the GTR substitution model and partitioned so codon positions 1 and 2 shared a substitution model while codon position 3 had its own substitution model. Amongst site rate heterogeneity was accounted for using gamma rate heterogeneity with 4 gamma classes. We applied a coalescent Bayesian skyline tree prior. We used the strict clock and relaxed lognormal clock models to assess variation in substitution rate within each dataset. The priors on the substitution rate were chosen to encompass the 95% highest probability density (HPD) of previously published estimates (Bok et al., 2009; Siebenga et al., 2010). We employed a lognormal prior on the capsid substitution rate with mean 4.3×10^{-3} substitutions/site/year and standard deviation 0.1 (Bok et al., 2009), while for the RdRp we used a lognormal prior on the substitution rate with mean 4.32×10^{-3} substitutions/site/year and standard deviation 0.1 (Siebenga et al., 2010). At the time of our analyses, there was no previously published estimate of the VP2 substitution rate across the entire GII.4 clade. We therefore applied the same substitution

rate prior to VP2 as the capsid. We performed three replicate runs with different starting parameters for each dataset and clock model and ran until convergence, as assessed using Tracer v1.5. We combined the three replicate runs with removal of suitable burnin using LogCombiner v2.2.1. Maximum clade credibility (MCC) trees were obtained with TreeAnnotator v2.2.1. The MCC trees obtained with BEAST consistently exhibited the same overall topology as the maximum likelihood trees reconstructed using RAxML.

3.3.5 Calculation of parameter estimates and 95% confidence intervals

All estimates of divergence dates, branch lengths and recombination dates were calculated by combining the complete posterior distribution of trees from each of the subsampled datasets. We inferred the date of divergence between each pair of strains by calculating the date of the most recent common ancestor of the two strains in each tree within the posterior distribution and then calculating the mean and 95% HPD of this distribution. We calculated the date at which a recombination event occurred by identifying the branch in the tree along which the recombination event occurred in each tree in the posterior distribution. We combined the dates of the start and end of this branch into a distribution of times for that branch and calculated the mean and 95% HPD of this distribution.

3.3.6 Acquisition of strain-specific datasets

We examined the evolutionary dynamics of the New Orleans 2009 and Sydney 2012 strains using datasets containing all of the P2 domain sequences with a known collection date available for the strain as of 30/10/2015. All analyses were carried out on the New Orleans 2009 dataset and the Sydney 2012 dataset independently. The datasets for each strain were aligned at the amino acid level using MUSCLE (Edgar, 2008). Each of the datasets exhibited a significant ($p < 0.001$) temporal signal, as assessed using a maximum likelihood tree as described above. The evolutionary dynamics of the P2 domain datasets

were reconstructed with BEAST v2.2.1. We applied the HKY substitution model with gamma rate heterogeneity with four gamma classes. We used the strict clock and relaxed lognormal clock models to test for variation in the substitution rate within the strain. In each case, there was strong support (log10 Bayes factor > 100) to reject the strict clock model in favour of the relaxed lognormal clock model. We employed a lognormal prior on the substitution rate with mean 6.83×10^{-3} and standard deviation 0.1 to accomodate the mean and the 95% HPD of our estimate of the substitution rate of the complete GII.4 capsid clade. Triplicate runs were carried out for each dataset and were run until convergence, as assessed with Tracer v1.5. Runs were combined with removal of suitable burnin using LogCombiner v2.2.1. Maximum clade credibility trees were obtained using TreeAnnotator v2.2.1. Bayesian skyline plots were reconstructed using Tracer v1.5. We used the date at which the Bayesian skyline plot exhibits a large increase in relative genetic diversity as the date of the onset of the pandemic. To determine the number of lineages present at the onset of the pandemic, we calculated the number of lineages present at this point in time in each tree in the posterior distribution.

3.3.7 Identification of pre-pandemic and pre-epidemic sequences

We defined pre-pandemic/pre-epidemic sequences as sequences that cluster with a pandemic/epidemic strain but were collected prior to the year of onset of the pandemic/epidemic. We initially obtained all GII.4 capsid sequences present on GenBank containing more than 400 nucleotides and genotyped each sequence using the norovirus genotyping tool (Kroneman et al., 2011). We identified 50 sequences that were associated with a collection date that was earlier than the year in which the respective pandemic/epidemic strain emerged. We assessed whether each of these putative pre-pandemic/pre-epidemic sequence had accumulated close to the expected amount of nucleotide change given its reported collection date by reconstructing a maximum likelihood tree of the P2 domain with RAxML (Stamatakis, 2014) as described above and assessed this tree using TempEst v1.5 (Rambaut et al., 2016). We estimated the collection date of each putative pre-pandemic/pre-epidemic sequence using BEAST v2.4.2 (Bouckaert et al., 2014). We used the capsid subsample 1 dataset assembled in section 3.3.4

and added to this dataset the putative pre-pandemic/pre-epidemic sequences in batches of five. The collection dates of the capsid subsample 1 dataset were fixed to the most accurate date possible, while we employed a uniform prior distribution on the collection date of the putative pre-pandemic/pre-epidemic sequences with minimum 1974.5 and maximum 2015.446575, the earliest and latest collection dates within the capsid subsample 1 dataset. We then estimated the collection date of each of these sequences as part of the MCMC chain. We used the SRD09 model of nucleotide substitution and a relaxed lognormal clock model. We assumed a lognormal prior on the substitution rate with mean 6.83×10^{-3} and standard deviation 0.1, chosen to encompass the mean and 95% HPD estimates of the GII.4 capsid substitution rate estimated in the BEAST runs described in section 3.3.4 (Table 3.2). We carried out triplicate runs and run until convergence as assessed using Tracer v1.6. Replicate runs were combined using LogCombiner v2.2.1.

3.4 Results

3.4.1 The GII.4 capsid emerged by the 1940's

We collected datasets for the GII.4 capsid, VP2 and associated RdRps and screened these datasets for potential recombinants (Table S3.1). The GII.4 capsid, VP2 and associated RdRps each exhibit a significant correlation between root-to-tip distance and sampling date, indicating there has been an accumulation of nucleotide change through time in each genomic region (Figure 3.1). We therefore reconstructed the evolutionary dynamics of each of the genomic regions using BEAST (Figure 3.1, Table 3.2). There is strong support to reject the strict clock model in favour of the relaxed lognormal clock model in each genomic region (Table 3.2). The substitution rate is very similar for the RdRp, capsid and VP2 (Table 3.2). Previous estimates have placed the common ancestor of the GII.4 capsid in the 1960s (Bok et al., 2009). In our dataset, the GII.4 capsid common ancestor occurred in the 1940s, due to the inclusion of a sequence collected from Malaysia in the 1970s that branches from the root of the GII.4 clade (Figure 3.1) and was not included in the previous study. The proposed GII.4 root in the previous study falls

Genome region	Log10 BF rejecting strict clock	Substitution rate, x10^{-3} substitutions/site/year (95% HPD)	Root date (95% HPD)	GII.P4 RdRp ancestor date (95% HPD)
RdRp	54.89	6.67 (5.84,7.53)	1941 (1922,1958)	1979 (1975,1982)
Capsid	122.97	6.83 (6.03,7.68)	1943 (1919,1965)	N/A
VP2	44.23	6.11 (5.35,6.93)	1949 (1934,1963)	N/A

Table 3.2: Summary of Bayesian MCMC results. Mean values and 95% HPD intervals were calculated by combining the complete posterior distributions for each parameter of interest from each of the subsampled datasets. The log10 Bayes factor (BF) is the support for rejecting the strict clock model in favour of the relaxed lognormal clock model.

within our phylogeny at a similar date (node P in Figure 3.1). Given the paucity of GII.4 sequences from and prior to the 1970s, it is possible that the GII.4 capsid common ancestor occurred earlier than our estimate. The GII.4 viruses collected in the 1970s are not monophyletic in the capsid tree. However, each of these samples has the GII.P1 RdRp, suggesting that the GII.4 capsid common ancestor was found with the GII.P1 RdRp. The GII.P4 RdRp found with the majority of more recent GII.4 capsids arose by the late 1970s (Figure 3.1, Table 3.2).

3.4.2 GII.4 pandemic strains are present for years prior to causing the pandemic

Each of the pandemic and epidemic GII.4 strains was present for years prior to causing their respective pandemic or epidemic, as evidenced by the occurrence of the common ancestor of each strain years prior to pandemic/epidemic spread (Table 3.3). However, the common ancestor is the ancestor of the subset of diversity represented in our datasets and is unlikely to represent the true common ancestor of the strain. Indeed, the branch leading to the common ancestor of each of the pandemic GII.4 strains typically accounts for several years of unsampled diversity (Figure 3.1, Table 3.3) and the true common ancestor of the strain could have occurred at any point along this branch. If a GII.4 strain evolved from one of the preceding pandemic strains, the new strain would diverge from its predecessor during the time at which that previous strain was circulating. However, the GII.4 strains exhibit deep divergence times, years prior to the emergence of either of the strains (Figure 3.2, Table 3.4), demonstrating that (with the possible exception of emergence

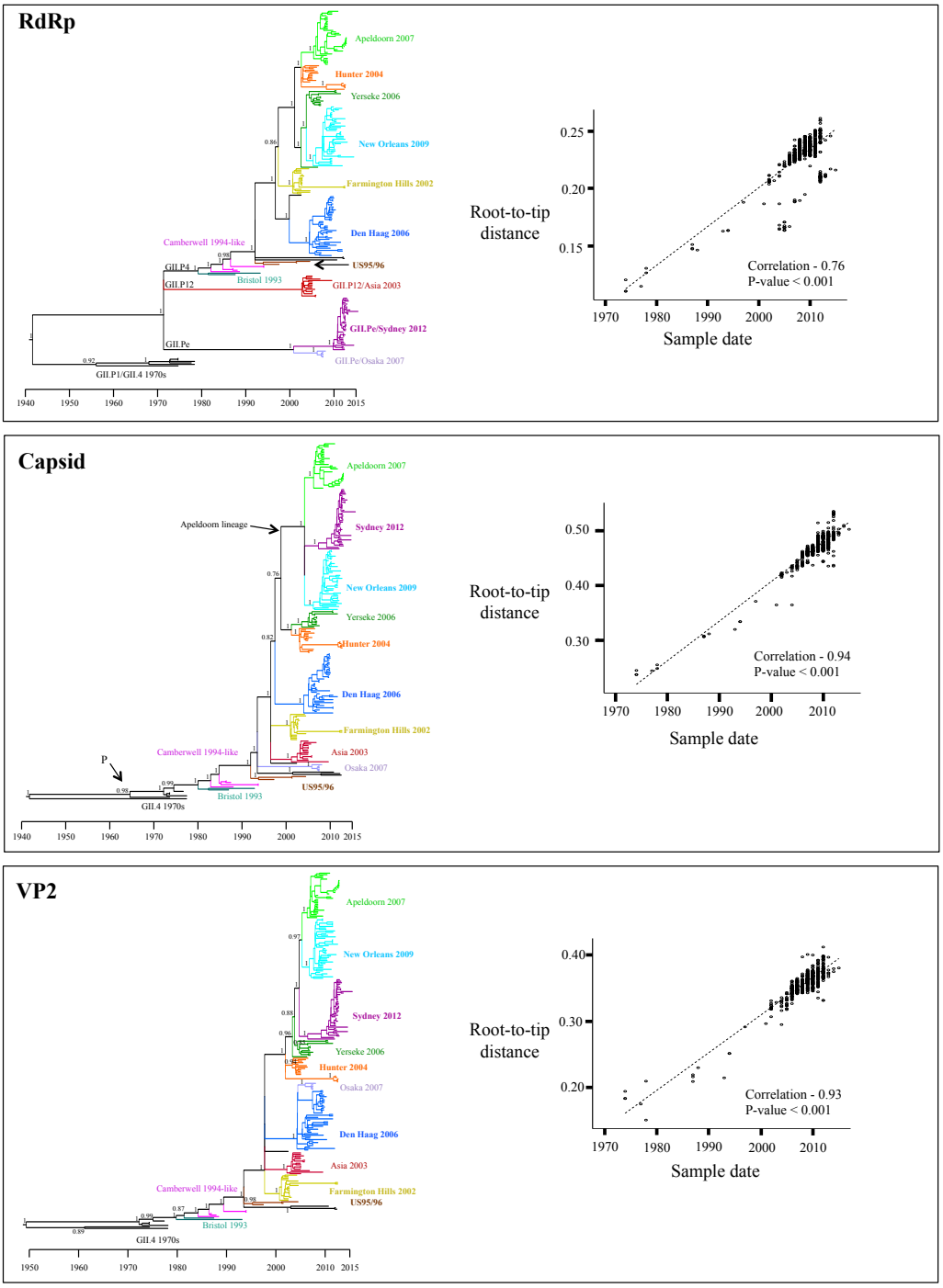


Figure 3.1: Temporal evolutionary history of the GII.4 capsid, VP2 and associated RdRps. Maximum clade credibility (MCC) trees are shown for the RdRp, capsid and VP2 as reconstructed using BEAST. Poorly supported (posterior support < 0.75) trunk nodes have been collapsed. Branches within each tree are coloured by the corresponding strain, with the colours matching the strain label. Pandemic strains are labelled in bold font. Posterior supports are shown on trunk nodes and strain root nodes. The posterior support on the US95/96 strain root node in the RdRp tree is 0.76. Node P within the capsid tree corresponds to the root node of the GII.4 capsid clade in a previous analysis (Bok et al., 2009). The date of node P is very similar to the root date estimated in this previous analysis. The correlation between root-to-tip distance and collection date is shown for sequences within each genomic region. The p-value on the R^2 correlation was calculated using a bootstrapping approach whereby the sampling dates were randomised and the R^2 correlation recalculated 1000 times. The significant correlation enabled reconstruction of the temporal evolutionary history of each genomic region.

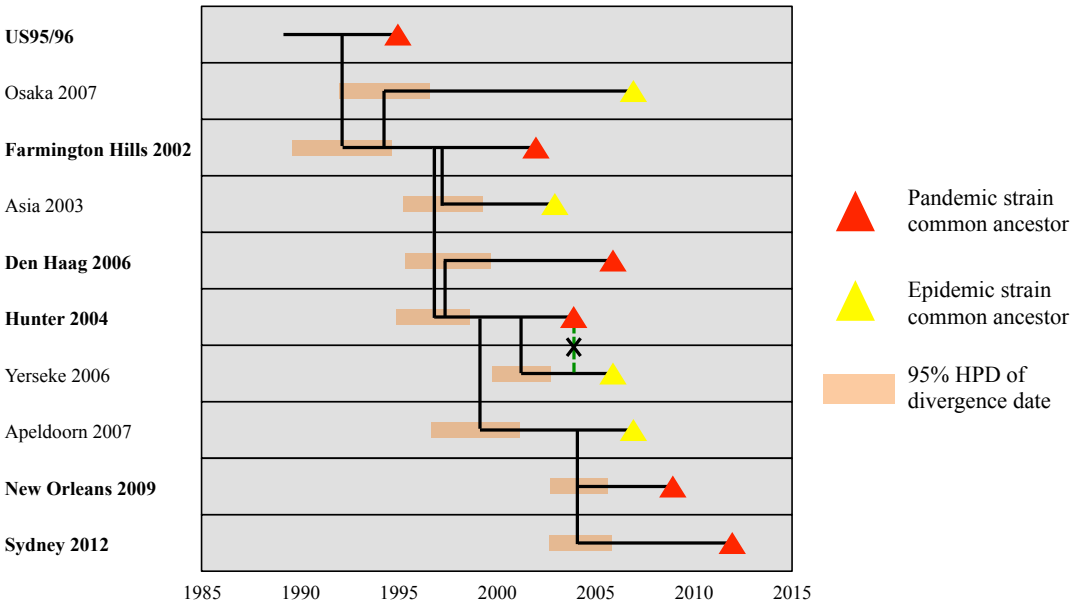


Figure 3.2: Comparison of divergence and emergence times. Shown is a representation of the capsid tree in Figure 3.1. The black vertical line occurs at the mean time of divergence and the orange shaded area represents the 95% HPD on the divergence time. The triangle represents the start of the year in which the strain emerged either pandemically (red) or epidemically (yellow). If a GII.4 strain evolved from a preceding strain, the divergence would occur after the emergence (triangle) of the preceding strain, as illustrated by the (unobserved) dashed green line. In all cases, divergence occurs earlier: Hunter 2004 diverged from Farmington Hills 2002 and Asia 2003 before either the Farmington Hills 2002 pandemic or the Asia 2003 epidemic, Den Haag 2006 diverged from Hunter 2004 before the Hunter 2004 pandemic, etc.

from US95/96 for which we have very few sequences) none of the pandemic or epidemic GII.4 strains evolved from a previous pandemic or epidemic strain. Rather, GII.4 strains diverge from and coexist with other pandemic and pre-pandemic strains for years prior to causing the pandemic in the human population. The fact that each of the strains exists for years before emergence, while only rarely being sampled, suggests that the strains circulate within one or more populations that is hidden from current surveillance.

GII.4 Strain	Common ancestor date			Years of unsampled diversity leading to strain common ancestor		
	RdRp (95% HPD)	Capsid (95% HPD)	VP2 (95% HPD)	RdRp (95% HPD)	Capsid (95% HPD)	VP2 (95% HPD)
US95/96	December 1993 (November 1991 - January 1996)	December 1993 (June 1991 - April 1996)	April 1995 (September 1993 - September 1996)	1.2 (0.0, 2.9)	1.8 (0.1, 4.1)	1.4 (0.2, 2.9)
Farmington Hills 2002	July 2000 (August 1999 - April 2001)	December 2000 (February 2000 - September 2001)	September 2000 (October 1999 - July 2001	3.2 (1.5, 5.1)	3.2 (1.4, 5.1)	2.5 (1.0, 4.1)
Asia 2003	August 2002 (June 2001 - August 2003)	February 2001 (May 1999 - November 2002	March 2002 (November 2000 - April 2003	29.1 (19.0, 40.2)	3.7 (1.6, 5.9)	3.8 (1.8, 5.8)
Hunter 2004	September 2002 (August 2001 - August 2002)	December 2002 (March 2002 - September 2003)	April 2002 (September 2000 - August 2003)	0.9 (0.0, 2.2)	1.7 (0.6, 3.0)	1.3 (0.3, 2.4)
Yerseke 2006	June 2002 (January 2001 - October 2003)	July 2003 (April 2002 - October 2004)	April 2003 (April 2002 - April 2004)	1.5 (0.3, 2.8)	2.3 (0.8, 3.8)	1.0 (0.0, 2.1)
Den Haag 2006	March 2004 (February 2003 - April 2005)	February 2004 (October 2002 - April 2005)	January 2004 (October 2002 - April 2005	4.4 (2.6, 6.4)	5.6 (2.7, 8.4)	4.3 (2.2, 6.5)
Osaka 2007	February 2006 (January 2005 - January 2007)	September 2005 (June 2004 - November 2006)	December 2005 (November 2004 - November 2006)	5.4 (1.5, 10.6)	10.2 (4.8, 13.9)	1.3 (0.3, 2.5)
Apeldoorn 2007	April 2005 (April 2004 - March 2006)	January 2006 (December 2004 - December 2006)	April 2006 (June 2005 - January 2007)	2.5 (1.2, 3.7)	1.9 (0.6, 3.3)	1.0 (0.3, 1.8)
New Orleans 2009	April 2005 (December 2003 - July 2006)	December 2005 (October 2004 - February 2007)	October 2006 (November 2005 - August 2007)	1.5 (0.4, 2.6)	1.7 (0.5, 3.0)	1.5 (0.6, 2.5)
Sydney 2012	October 2009 (August 2008 - September 2010)	February 2007 (April 2005 - October 2008)	May 2006 (August 2004 - April 2008	9.0 (5.0, 14.4)	2.9 (1.1, 4.8)	1.7 (0.2, 3.4)

Table 3.3: Strain common ancestor dates and unsampled diversity. The common ancestor date of each GII.4 strain is shown for the RdRp, capsid and VP2. The number of years of unsampled diversity is the branch length leading to the common ancestor of each GII.4 strain. The common ancestor of the Sydney 2012 RdRp is the common ancestor of the GII.Pe RdRps found with the Sydney 2012 capsid (Figure 3.1 RdRp). Asia 2003 has a GII.P12 RdRp, which is also found with capsids from multiple other genotypes. As only GII.P12 RdRps associated with the GII.4 capsid were included here, the years of unsampled diversity leading to the Asia 2003 RdRp ancestor is larger than would be obtained if RdRps associated with other capsid genotypes were also included. 95% HPD intervals were calculated by combining the posterior distributions for each subsampled dataset. Pandemic strains are labelled in bold font.

GII.4 strain	**US95/96**	**Farmington Hills 2002**	Asia 2003	**Hunter 2004**	Yerseke 2006	**Den Haag 2006**	Osaka 2007	Apeldoorn 2007	**New Orleans 2009**	**Sydney 2012**
US95/96	-	1992 (1989, 1994)	1992 (1989, 1994)	1992 (1989 1994)	1992 (1989, 1994)	1992 (1989, 1994)	1992 (1989, 1994)	1992 (1989, 1994)	1992 (1989, 1994)	1992 (1989, 1994
Farmington Hills 2002	-	-	1997 (1995, 1999)	1996 (1994, 1998)	1996 (1994, 1998)	1997 (1994, 1999)	1994 (1991, 1996)	1996 (1994, 1998)	1996,1994, 1998)	1996 (1994, 1998)
Asia 2003	-	-	-	1996 (1994, 1998)	1996 (1994, 1998)	1996 (1994, 1998)	1994 (1991, 1996)	1996 (1994, 1998)	1996 (1994, 1998)	1996 (1994, 1998)
Hunter 2004	-	-	-	-	2001 (1999, 2002)	1997 (1995, 1999)	1994 (1991, 1996)	1999 (1996, 2001)	1999 (1996, 2001)	1999 (1996, 2001)
Yerseke 2006	-	-	-	-	-	1997 (1995, 1999)	1994 (1991, 1996)	1999 (1996, 2001)	1999 (1996, 2001)	1999 (1996, 2001)
Den Haag 2006	-	-	-	-	-	-	1994 (1991, 1996)	1997 (1995, 2000)	1997 (1995, 2000)	1997 (1995, 2000)
Osaka 2007	-	-	-	-	-	-	-	1994 (1991, 1996)	1994 (1991, 1996)	1994 (1991, 1996)
Apeldoorn 2007	-	-	-	-	-	-	-	-	2004 (2002, 2005)	2003 (2002, 2005)
New Orleans 2009	-	-	-	-	-	-	-	-	-	2004 (2002, 2005)
Sydney 2012	-	-	-	-	-	-	-	-	-	-

Table 3.4: Summary of strain divergence times. The mean divergence time for each pair of strains in the capsid posterior distribution of trees is shown. The 95% HPD of the divergence time is shown in parentheses. Pandemic strains are labelled in bold font.

3.4.3 Pandemic strains diverge into multiple lineages over several years prior to emergence

We next reconstructed the evolutionary dynamics of the two most recent pandemic strains, New Orleans 2009 and Sydney 2012, using a larger dataset of P2 domain sequences. New Orleans 2009 and Sydney 2012 underwent a large increase in relative genetic diversity in 2009 and 2012, respectively, coinciding with the time at which they emerged pandemically and replaced the preceding pandemic strain (Figure 3.3). However, by the time of the increase in relative genetic diversity, each strain had already diverged into multiple lineages; approximately 67 lineages (95% HPD 41-100 lineages) in the case of New Orleans 2009 and approximately 88 lineages (95% HPD 59-113 lineages) in the case of Sydney 2012 (Figure 3.3). Each of the other pandemic strains also exhibits this pre-divergence into multiple lineages prior to pandemic emergence, as evidenced by the occurrence of the strain common ancestor years prior to pandemic onset. However, the interval between common ancestor and emergence is largest for the New Orleans 2009 and Sydney 2012 pandemics than for the other pandemic strains, a finding that likely reflects more thorough surveillance from a greater number of countries in recent times. This suggests that the interval between common ancestor and pandemic or epidemic emergence for other strains is likely to be underestimated. These observations suggest that considerable and long-standing diversity of pandemic and epidemic strains exists within the hidden populations prior to their pandemic or epidemic emergence.

3.4.4 Recombination in the GII.4 genotype

There are topological differences between the RdRp, capsid and VP2 trees in well supported regions, indicating substantial recombination close to ORF boundaries (Figures 3.1, 3.4). In addition to the previously reported acquisition of a Den Haag 2006-like VP2 by the Osaka 2007 capsid (Eden et al., 2013), we also find that the Apeldoorn lineage acquired a Yerseke 2006-like VP2 in 2003, prior to diverging into the Apeldoorn 2007, New Orleans 2009 and Sydney 2012 strains (Figures 3.1, S3.2, Table 3.5). Multiple recombination events have occurred between the RdRp and capsid (Figure 3.4). In agreement

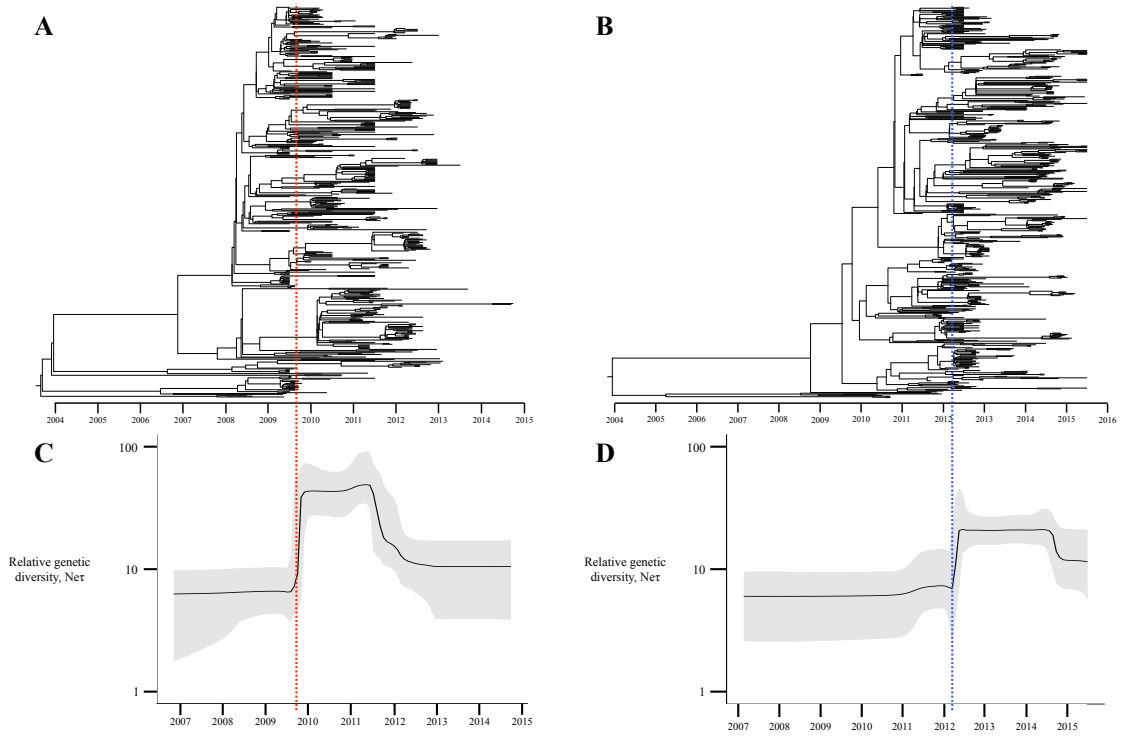


Figure 3.3: Evolutionary dynamics of New Orleans 2009 and Sydney 2012. Summary of phylodynamic analysis of the P2 domains of New Orleans 2009 (**A** and **C**) and Sydney 2012 (**B** and **D**). (**A** and **B**) MCC tree of the New Orleans 2009 and Sydney 2012 P2 domain datasets as reconstructed using BEAST. (**C** and **D**) Bayesian skyline plots showing the population dynamics of New Orleans 2009 and Sydney 2012 through time. The solid black line is the median value and the shaded grey area is the 95% HPD. An increase in Neτ indicates an increase in the relative number of viral infections and therefore in this case indicates the onset of the pandemic. The red and blue vertical dashed lines represent the time of the large increase in Neτ for New Orleans 2009 and Sydney 2012, respectively, by which time the strains had diverged into a large number of lineages.

with previous results, we find that Asia 2003 acquired a GII.P12 RdRp and Osaka 2007 acquired a GII.Pe RdRp (Eden et al., 2013). A common ancestor of the New Orleans 2009 and Sydney 2012 (and possibly the Apeldoorn 2007) strains acquired a Yerseke 2006-like RdRp in 2004 (95% HPD 2003-2006) (Figure S3.2, Table 3.5). Multiple recombination events are required to explain the acquisition of RdRp and VP2 regions by the Apeldoorn 2007 lineage capsid (consisting of the Apeldoorn 2007, New Orleans 2009 and Sydney 2012 strains) (Figure S3.2). The Sydney 2012 capsid sequences in our dataset are found with a New Orleans 2009-like RdRp and a GII.Pe RdRp. At least three independent recombination events are required to explain the distribution of the GII.Pe RdRps in the Sydney 2012 clade in the capsid tree (Figure S3.3). The clustering of the GII.Pe RdRps in the RdRp and capsid trees indicates the cocirculation of the Sydney 2012 viruses with

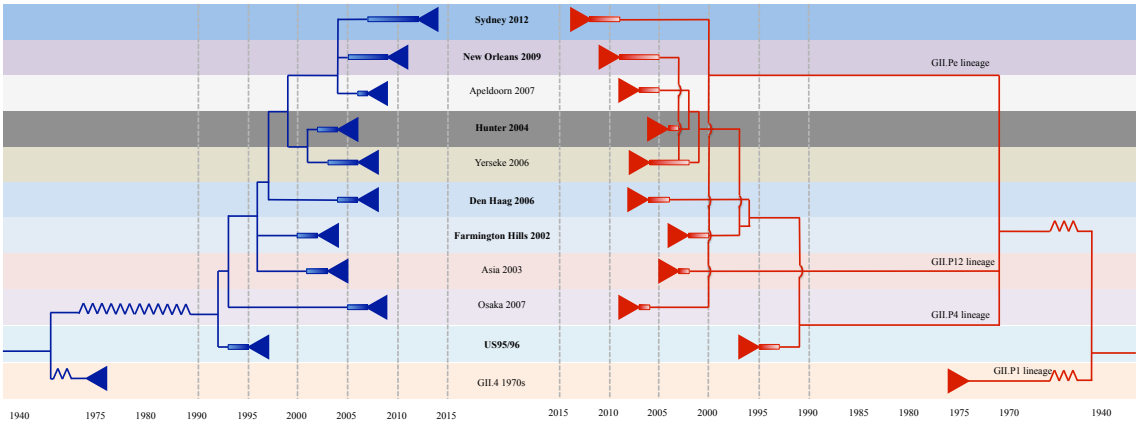


Figure 3.4: Comparison of the temporal evolutionary history of the GII.4 capsid and associated RdRps. Representations of the capsid (blue) and RdRp (red) trees from Figure 3.1 are shown. Nodes with posterior support less than 0.75 have been collapsed. For each strain, the boxed area represents the time from the common ancestor of the strain to the start of the year in which the strain emerged pandemically or epidemically, with the triangle representing time after the strain has emerged. The pandemic GII.4 strains are labelled in bold font.

Capsid strain	RdRp or VP2 acquired	Time of recombination event (95% HPD)
Asia 2003	GII.P12 RdRp	March 1999 (March 1996-October 2001)
Osaka 2007	GII.Pe RdRp	June 2000 (October 1994-January 2006)
Osaka 2007	Den Haag 2006-like VP2	March 2005 (December 2003-August 2006)
Apeldoorn lineage	Yerseke 2006-like VP2	March 2004 (February 2003-April 2005)
Apeldoorn lineage	Yerseke 2006-like RdRp	July 2004 (January 2003-November 2005)
Sydney 2012	GII.Pe RdRp	May 2010 (July 2009-February 2011) January 2012 (July 2011-May 2012) February 2012 (August 2011-July 2012)

Table 3.5: Summary of recombination events in the GII.4 lineage. Shown are the dates at which the capsid gene acquired a new RdRp or VP2. Recombination events were initially identified on the basis of well-supported topological differences between the capsid tree and the RdRp or VP2 tree. The date of the recombination event was calculated from the complete posterior distribution of trees in the corresponding dataset by identifying the branch along which the recombination event occurred in each tree. Sydney 2012 acquired the GII.Pe RdRp in at least three independent recombination events. Additional recombination events likely occurred within the Apeldoorn lineage. Further information on recombination within the Apeldoorn lineage is shown in Figures S3.2 and S3.3.

both GII.P4 New Orleans 2009-like and GII.Pe RdRps, consistent with previous results (Wong et al., 2013). Importantly, each of the recombination events where the two contributing strains can be identified occurred years prior to the emergence of either of the contributing strains (Table 3.5).

3.4.5 Identification of pre-pandemic GII.4 viruses

We next identified publically available pre-pandemic/pre-epidemic sequences, defined here as sequences that cluster with a GII.4 strain but were collected prior to the year of onset of the respective pandemic or epidemic. We identified 50 putative pre-pandemic/pre-epidemic sequences that were genotyped as either Farmington Hills 2002, Hunter 2004, Osaka 2007, New Orleans 2009 or Sydney 2012 (Tables 3.6, 3.7). To validate whether the reported collection dates for these sequences were plausible, we estimated the collection date of each sequence using BEAST (Figure S3.4, Tables 3.6, 3.7). The posterior estimate of the collection date overlapped with the reported collection date for 31 sequences (Figures 3.5, S3.4, Table 3.6). Each of these sequences also exhibits close to the expected accumulation of nucelotide change given their reported collection date (Figure S3.5). Our analysis therefore supports, but does not confirm, the reported collection date for these sequences and suggests that these sequences are true pre-pandemic/pre-epidemic sequences. We therefore find evidence for pre-pandemic sequences from the Farmington Hills 2002, Hunter 2004 (Sdiri-Loulizi et al., 2009), New Orleans 2009 (Eden et al., 2014) and Sydney 2012 (Eden et al., 2014) strains as well as pre-epidemic sequences from the Osaka 2007 strain (Figure 3.5). Of the 19 sequences where the reported and estimated collection dates do not overlap, 18 have an earlier reported collection date than that estimated by our analysis (Table 3.7). Possible reasons for this include mis-reporting of the collection date (most of these sequences are not associated with a publication and it is therefore not possible to verify the collection date in primary literature) or sample contamination or mis-labelling.

3.5 Discussion

The suggestion that each of the pandemic GII.4 strains evolved from one of the previous pandemic or epidemic strains (Siebenga et al., 2007; van Beek et al., 2013) has recently come into question with the demonstration that New Orleans 2009 and Sydney 2012 diverged from a common ancestor close to the start of the Apeldoorn 2007 epidemic (Eden et al., 2014). We demonstrate here that this finding can be generalised. None of

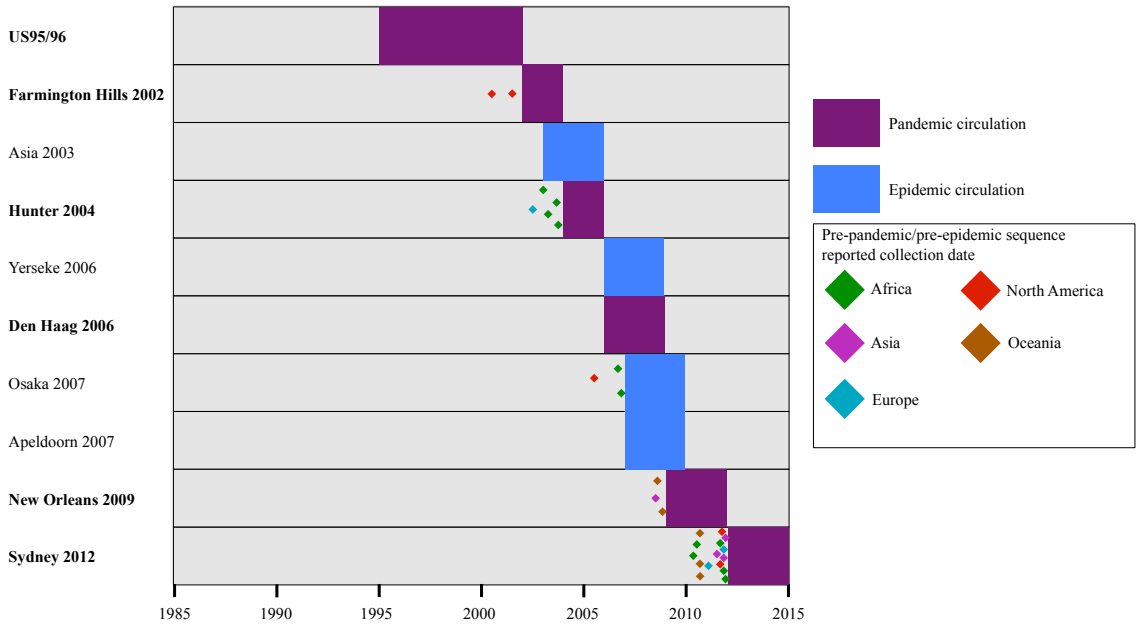


Figure 3.5: Pre-pandemic and pre-epidemic GII.4 sequences. We collected and genotyped all available norovirus sequences from GenBank and identified 50 GII.4 sequences with a reported collection date earlier than the year of onset of the pandemic/epidemic caused by the strain to which the sequence was genotyped. We estimated the collection date of each of these sequences using BEAST and found 31 sequences where the estimated collection date overlapped with the reported collection date; this analysis supports, but does not confirm, the reported collection date for these sequences and suggests that these sequences are true pre-pandemic/pre-epidemic sequences. The reported collection date for each of these sequences is shown here as a diamond coloured by the continent on which the sequence was collected. The putative pre-pandemic Sydney 2012 sequences LC005721.1, LC005722.1, LC005723.1 and LC005724.1 were all collected in December 2011 in the same study (Kumazaki and Usuku, 2015) (Table 3.6); a single diamond is shown for these four sequences for clarity. The period of pandemic (purple) or epidemic (blue) circulation is shown for each GII.4 strain. Further information on the putative pre-pandemic and pre-epidemic sequences is shown in Figures S3.4, S3.5 and Table 3.6.

the pandemic GII.4 strains evolved from a previous pandemic or epidemic strain (with the possible exception of evolving from US95/96); once a GII.4 strain has emerged pandemically or epidemically, it does not give rise to any future strains. Instead, the GII.4 phylogenetic trees indicate that future pandemic strains circulate and evolve within one or more poorly sampled reservoirs containing substantial genetic diversity. For example, during the late 1990s when US95/96 was the pandemic strain, we estimate that there were at least six other distinct lineages of GII.4 capsid present (Farmington Hills 2002, Asia 2003, Den Haag 2006, Osaka 2007, the Hunter lineage (leading to Hunter 2004 and Yerseke 2006) and the Apeldoorn lineage (leading to Apeldoorn 2007, New Orleans 2009

Accession number	Strain	Capsid region present	Country of collection	Reported collection date	Estimated collection date (95% HPD)
EU078408.1	Farmington Hills 2002	Complete	USA	2001	2000 (1998-2003)
GU937448.1	Farmington Hills 2002	P domain	USA	2000	2000 (1998-2003)
EU916961.1	Hunter 2004	Complete	Tunisia	11-January-2003	2004 (2003-2006)
EU916960.1	Hunter 2004	Complete	Tunisia	11-September-2003	2003 (2002-2005)
EU916959.1	Hunter 2004	Complete	Tunisia	05-April-2003	2004 (2003-2006)
EU916957.1	Hunter 2004	Complete	Tunisia	17-October-2003	2004 (2003-2006)
EU876890.1	Hunter 2004	Complete	France	2002	2004 (2002-2006)
GQ845367.2	New Orleans 2009	Complete	Australia	November-2008	2008 (2006-2010)
GQ845345.2	New Orleans 2009	Complete	Australia	August-2008	2008 (2006-2009)
KR131773.1	New Orleans 2009	Complete	India	2008	2009 (2008-2010)
KJ735099.1	Sydney 2012	Missing first 25 AA	Morocco	14-September-2011	2011 (2009-2012)
KF509947.2	Sydney 2012	Complete	Canada	September-2011	2010 (2007-2012)
AB972499.1	Sydney 2012	Complete	Japan	2011	2012 (2011-2013)
LC005724.1	Sydney 2012	Complete	Japan	December-2011	2012 (2010-2014)
LC005723.1	Sydney 2012	Complete	Japan	December-2011	2011 (2010-2013)
LC005722.1	Sydney 2012	Complete	Japan	December-2011	2011 (2010-2013)
LC005721.1	Sydney 2012	Complete	Japan	December-2011	2011 (2010-2012)
LC005720.1	Sydney 2012	Complete	Japan	November-2011	2012 (2010-2015)
KR904236.1	Sydney 2012	Complete	South Africa	19-December-2011	2009 (2006-2012)
KR904215.1	Sydney 2012	Complete	South Africa	06-July-2010	2009 (2006-2011)
KR904214.1	Sydney 2012	Complete	South Africa	19-May-2010	2009 (2006-2012)
KC962462.3	Sydney 2012	Complete	South Africa	November-2011	2010 (2007-2013)
KF870711.1	Sydney 2012	AA 299-end	Spain	Febuary-2011	2010 (2008-2015)
KF668567.1	Sydney 2012	Complete	Italy	01-November-2011	2011 (2010-2013)
KF060124.1	Sydney 2012	Complete	New Zealand	September-2010	2009 (2006-2012)
KF060123.1	Sydney 2012	Complete	New Zealand	September-2010	2009 (2006-2011)
KF060122.1	Sydney 2012	Complete	New Zealand	September-2010	2009 (2006-2012)
KX354057.1	Sydney 2012	Complete	USA	26-October-2011	2012 (2010-2015)
FJ411171.1	Osaka 2007	Complete	USA	2005	2003 (1999-2007)
EU876884.1	Osaka 2007	Complete	Egypt	November-2006	2004 (2000-2008)
EU876882.1	Osaka 2007	Complete	Egypt	September-2006	2004 (2000-2008)

Table 3.6: Summary of pre-pandemic and pre-epidemic GII.4 sequences. We collected and genotyped all available norovirus sequences and identified 50 GII.4 sequences with a reported collection date earlier than the year of onset of the pandemic/epidemic caused by the strain to which the sequence was genotyped. We estimated the collection date of each of these 50 sequences using BEAST. The 31 sequences where the 95% HPD of the estimated collection date overlaps with the reported collection date were taken to be true pre-pandemic/pre-epidemic sequences. These sequences are summarised here. The strain is the strain as genotyped by the norovirus genotyping tool (Kroneman et al., 2011); all assignments were supported by phylogenetic clustering in a nucleotide maximum likelihood phylogenetic tree. Note that while the estimated collection date is given to the year, we compared the estimated and reported collection dates to the most precise value possible; therefore where the reported collection date was given to the nearest day or month, the 95% HPD of the estimated collection date overlaps with that day or month. AA - amino acid.

and Sydney 2012)) that had not yet been sampled. This high diversity present in reservoirs is also supported by recombination events between multiple GII.4 strains prior to their pandemic or epidemic emergence (Table 3.5). Interestingly, our results suggest that the majority of the diversity that has seeded pandemics and epidemics post-US95/96 had been generated by the year 2000. As discussed and demonstrated in the previous chapter, US95/96 was likely the first GII.4 pandemic and the onset of this pandemic likely coincided with GII.4 transitioning from one of a number of relatively low frequency genotypes to the dominant genotype in the human population (Donaldson et al., 2008, 2010; Siebenga et al., 2010). This first pandemic therefore likely resulted in a large population expansion for GII.4 noroviruses and appears to have resulted in the generation of

Accession number	Strain	Capsid region present	Country of collection	Reported collection date	Estimated collection date (95% HPD)
KU312299.1	Farmington Hills 2002	P2 domain	UK	1994	2002 (2001-2004)
EU916958.1	Hunter 2004	Missing first 3 AA	Tunisia	02-January-2003	2004 (2003-2005)[a]
KU182476.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182477.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2005)
KU182478.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182479.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182480.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182481.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182482.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU182483.1	Hunter 2004	Complete	Tunisia	2002	2004 (2003-2006)
KU312301.1	Hunter 2004	P2 domain	UK	1995	2004 (2002-2007)
KU312302.1	Hunter 2004	P2 domain	UK	1995	2004 (2002-2006)
KU312303.1	Hunter 2004	P2 domain	UK	1995	2004 (2002-2007)
AB972505.1	Sydney 2012	Complete	Japan	2011	2013 (2012-2014)
AB972504.1	Sydney 2012	Complete	Japan	2011	2013 (2012-2014)
AB972503.1	Sydney 2012	Complete	Japan	2011	2013 (2012-2015)
AB972502.1	Sydney 2012	Complete	Japan	2011	2013 (2012-2014)
KU312300.1	Sydney 2012	P2 domain	UK	1994	2013 (2011-2015)
FJ411172.1	Osaka 2007	Complete	USA	2006	2001 (1997-2004)

Table 3.7: Summary of putative pre-pandemic and pre-epidemic sequences where the reported collection date was not supported by phylogenetic analyses. We collected and genotyped all available norovirus sequences and identified 50 GII.4 sequences with a reported collection date earlier than the year of onset of the pandemic/epidemic caused by the strain to which the sequence was genotyped. We estimated the collection date of each of these 50 sequences using BEAST. The 19 sequences where the 95% HPD of the estimated collection date does not overlap with the reported collection date were not taken to be pre-pandemic/pre-epidemic sequences. Rather, we suggest that the collection date for these sequences was likely mis-reported or that the samples were contaminated or mislabelled. These sequences are summarised here. While the estimated collection date of 18 of these sequences is later than the reported collection date, the estimated collection date of sequence FJ411172.1 is earlier than the reported collection date. The strain is the strain as genotyped by the norovirus genotyping tool (Kroneman et al., 2011); all assignments were supported by phylogenetic clustering in a nucleotide maximum likelihood phylogenetic tree. [a] The 95% HPD of the estimated collection date of sequence EU916958.1 does not overlap with the reported collection date to the level of the day. AA - amino acid.

a high level of diversity within a short space of time. Multiple studies have suggested that different pandemic strains may infect different sections of the human population via interactions with different HBGA receptors (Lindesmith et al., 2008; Donaldson et al., 2008). It will be interesting in the future to test the HBGA-binding profile of ancestral viruses from each of the lineages that diverged in the late 1990s to determine whether this rapid generation of diversity may have been driven by adaptation of different lineages to different HBGA-binding niches and therefore different parts of the human population.

From our observations, two questions arise. First, what drives a strain that has been circulating undetected for many years to spread rapidly, cause global outbreaks and replace the previous pandemic strain. Second, where are what are the reservoirs in which future pandemic strains spread and diversify but remain undetected? Our results suggest that the emergence of a new pandemic GII.4 strain occurs in three stages (Figure 3.6).

The predivergence into multiple lineages over years prior to the pandemic suggests that the mutations and/or recombination events required for pandemic spread (e.g. for evasion of herd immunity) occur years prior to pandemic spread, as otherwise the same change would need to occur independently many times. Therefore we predict that pandemic strains acquire all of the genetic changes necessary to cause a future pandemic while they are circulating within the unsampled reservoir population (stage 1). The pre-pandemic strains continue to diversify within these reservoirs (stage 2), resulting in multiple lineages containing the important pandemic-enabling changes. During this diversification, the strain may remain within a single geographical reservoir or may undergo limited spread to multiple geographically distinct reservoirs (Figure 3.6). Finally, the related lineages emerge concurrently, spreading rapidly to cause the new pandemic and replacing the previous pandemic strain (stage 3). While we have depicted these stages as discrete time steps, in reality the acquisition of changes and diversification is likely a continuous process, with many low level viruses being in stages 1 and 2 at any point in time.

While viral genetic changes are vital for pandemic emergence, their occurrence multiple years before pandemic spread indicates that these changes are not the proximate driver of the new pandemic, in constrast with previous suggestions (Eden et al., 2014). Rather, viral genetic changes enable the strain to emerge as a pandemic in the future. Given the very high number of GII.4 norovirus cases annually (Kirk et al., 2015; Pires et al., 2015; Lopman et al., 2016), it is very unlikely that stochastic factors could drive the emergence of a large number of lineages of one strain of one genotype. Similarly, it is hard to conceive of an event or set of environmental factors that could be responsible for this emergence. We therefore suggest that the emergence of a new pandemic GII.4 strain is most likely to be driven by a change in the host population. Given current evidence of the importance of herd immunity in GII.4 strain emergence (Lindesmith et al., 2008, 2012a; Debbink et al., 2013), we hypothesise that growing herd immunity against the previous pandemic strain may open a niche into which multiple lineages of the new strain can emerge. We therefore hypothesise that it is changes in the host immunity profile, rather than changes in the virus, that are the drivers of norovirus pandemic emergence. In the future, mathematical models incorporating viral antigenicity and transmission will be key to provide support for this hypothesis.

Multiple studies have suggested immunocompromised patients as a potential reser-

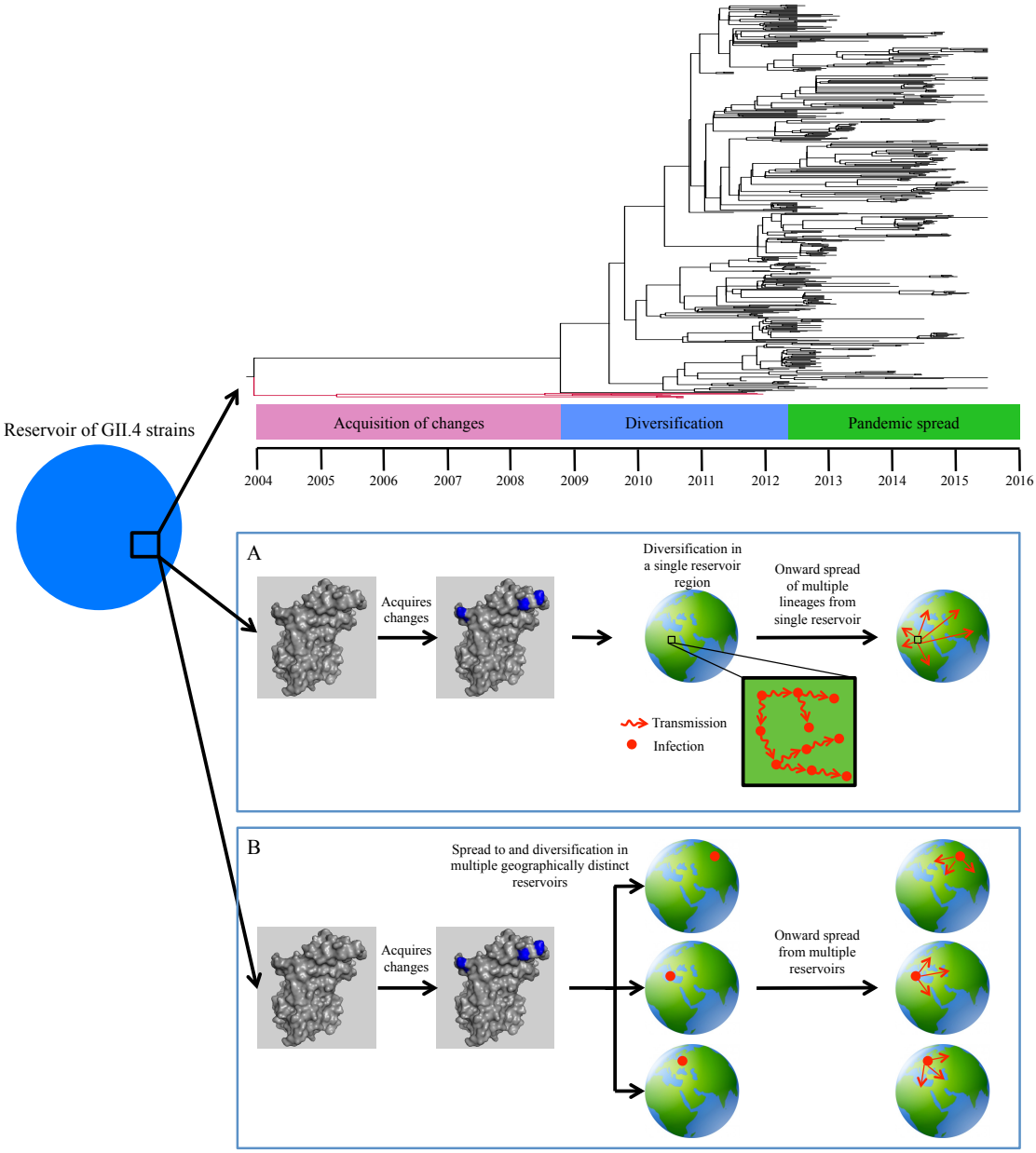


Figure 3.6: Three stage emergence of new pandemic GII.4 strains. The strain that will emerge as a pandemic is a component of an unsampled reservoir of GII.4 strains. The pandemic emergence of the new strain occurs in three stages. In the first stage, the strain acquires the genetic changes that will be essential for it to spread pandemically, with example changes shown in blue in the capsid P2 domain (PDB structure 2OBR). This acquisition occurs years prior to emergence, leaving the strain pre-adapted for future spread. In the second stage, the strain continues to diversify within the reservoir or reservoirs resulting in multiple lineages that share the pre-adaptation. The strain at this stage may be within a single reservoir region (**A**) or may undergo limited spread to multiple geographically distinct reservoir regions (**B**). Multiple lineages that share the pre-adaption then emerge concurrently in stage 3. Here, the three stages of strain emergence are shown with reference to the Sydney 2012 phylogenetic tree in Figure 3.3B. The red lineages that branch from the root of the Sydney 2012 tree did not persist and do not contain changes that have been suggested to be important for the pandemic emergence of Sydney 2012 (Debbink et al., 2013; Allen et al., 2014).

voir for GII.4 viruses due to the diverse viral population often found in these patients (Debbink et al., 2014; Vega et al., 2014b; Karst and Baric, 2015). There is also evidence of GII.4 infection of swine, cattle (Mattison et al., 2007) and dogs (Summa et al., 2012; Caddy et al., 2015), suggesting the possibility of an animal reservoir. Both of these options are possible. However, a single immunocompromised patient cannot be the reservoir for a new strain due to the diversification of the virus into multiple lineages prior to emergence and it is difficult to explain how immunocomprised patients could form the interconnected transmission network required for this strain diversification. Similarly, the emergence of the multiple lineages concurrently from an animal reservoir would require a large number of zoonotic transmissions at the onset of the pandemic; no such transmissions have yet been observed. Instead we propose that the future pandemic/epidemic strains are likely to circulate undetected in the human population in either the community or in undersampled geographical regions for at least several years prior to their emergence as a new pandemic, even if they initially arise in an immunocompromised patient or animal reservoir. While a recent study by Allen et al detected multiple pandemic GII.4 strains in community cases of sporadic diarrhoea in the UK 8-18 years before their emergence as a pandemic (Wheeler et al., 1999; Allen et al., 2016), our phylogenetic analysis did not support the reported collection dates for these sequences (Table 3.7). However, we do identify pre-pandemic sequences from the Farmington Hills 2002, Hunter 2004, New Orleans 2009 and Sydney 2012 pandemic strains, as well as pre-epidemic sequences from Osaka 2007 (Figure 3.5, Table 3.6). Interestingly, there are more pre-pandemic sequences from the Sydney 2012 pandemic strain compared with the other strains (Figure 3.5), correlating with increased and more widespread surveillance in recent times. It is interesting, however, that pre-pandemic samples of Hunter 2004 and Sydney 2012, as well as pre-epidemic sequences from Osaka 2007, have been identified in Africa, despite this region only being sparsely sampled (Sdiri-Loulizi et al., 2009; Kamel et al., 2009; Mans et al., 2016). Capsid sequences that fail to cluster within any of the major GII.4 strains have also been described, a feature that would be expected if Africa were a GII.4 reservoir (Mans et al., 2015, 2016). Importantly, current norovirus surveillance focuses on sequencing of outbreak strains, predominantly in hospital and institutional settings in middle and high income countries, a strategy that would miss community and low-income-country reservoirs. The majority of norovirus sampling is currently carried out

in outbreak settings, with few studies examining norovirus in the community (Inns et al., 2017). However, the studies of Infectious Intestinal Disease (IID1 and IID2) in the UK suggested norovirus is highly prevalent in community settings (Tam et al., 2012; Harris et al., 2017). It is therefore possible that the community could function as a reservoir of norovirus strains and could also be the location of the diversification we observe prior to emergence (Figure 3.6).

Our results suggest that the GII.4 strains that will become pandemic in the future are already circulating at low level, and may already be in their pandemic form. If regions of early circulation can be identified, it may be possible to identify strains currently at low level that are antigenically distinct from the currently dominant strain and therefore identify strains that have the potential to become pandemic in the future. Phylogeographic studies may identify the reservoir regions in which pre-pandemic circulation occurs and determine whether there is a single geographical reservoir, as has been demonstrated for for influenza A H3N2 (Bedford et al., 2015), or whether strains circulate widely prior to emergence. We employ a phylogeographic analysis to assess the potential for source regions in chapter 4. Recombination events occurring between multiple GII.4 strains years prior to the emergence of either strain (Table 3.5) argue for a single reservoir region as these recombination events require coinfection of individual patients with multiple low level strains. Alternatively, these viruses may come into contact during food or waterborne infections as these can be contaminated with multiple viruses (de Graaf et al., 2016).

Our hypothesis that new pandemics are not driven by viral genetic changes but by changes in host immunity represents a shift in our understanding of the drivers of new pandemics. We have discussed the importance of our results in developing surveillance strategies for the detection of new pandemic strains, but these results also have implications for the design of vaccines. Under the previous hypothesis where new pandemic strains typically evolve from previous pandemic strains, a vaccine targeting the current pandemic may help prevent future pandemics. If, however, the emergence of a new pandemic is driven by changes in host immunity, the targeting of a vaccine against the current strain may have the paradoxical effect of hastening the next pandemic. Therefore under our hypothesis it is essential that norovirus vaccines provide broad immunity against GII.4 noroviruses.

Chapter 4

Characterisation of global circulation and important substitutions for norovirus pandemics

4.1 Abstract

In the previous chapter, we suggested a three stage process of pandemic GII.4 strain emergence where the strain acquires the genetic changes that will be important for pandemic spread years prior to the pandemic and then undergoes diversification into multiple pre-adapted lineages that emerge concurrently to cause the pandemic. Here, we carry out phylogeographic analyses and demonstrate that the New Orleans 2009 and Sydney 2012 pandemic strains circulated widely and consistently during the diversification phase of strain emergence. We find no evidence for a single source region for pandemic GII.4 strains, but instead suggest that strains undergo low level worldwide spread over several years prior to the pandemic. At the onset of the pandemic strains therefore emerge from multiple geographical regions, strengthening our previous suggestions that strain emergence is driven by changes in host factors. We carry out an analysis of amino acid change to identify nonsynonymous substitutions that occurred leading to each of the pandemic GII.4 strains and resulted in a different amino acid residue(s) to that in the preceding pandemic strain, thereby identifying potential pandemic-enabling substitutions. We demonstrate that a high level of amino acid diversity exists within each of the pandemic GII.4 strains and that sites of diversity often coincide with epitope sites and/or HBGA-binding

sites. This suggests that viruses within a pandemic strain may exhibit subtly different phenotypes. In New Orleans 2009 and Sydney 2012, this diversity began to be accumulated prior to the pandemic emergence of the strains. These results strongly suggest that viruses within each pandemic strain share one or more phenotypic characteristics that are acquired by the strain root node and are vital to enable pandemic emergence. We extend our three stage model of strain emergence to suggest that strains move to a new region of antigenic space in the acquisition of changes phase and then undergo local movements within this new region of antigenic space in the diversification phase. We hypothesise that changes in host immunity enable each of the lineages within this region of antigenic space to emerge at the onset of the new pandemic.

4.2 Introduction

In the previous chapter, we provided evidence that each of the pandemic GII.4 strains circulated within one or more hidden reservoir populations for years prior to causing their respective pandemic. We demonstrated that each pandemic strain diverged into multiple lineages over years prior to pandemic onset, indicating that the genetic changes essential for pandemic spread are acquired years prior to the pandemic. Our results are supported by the detection of multiple pandemic strains in sporadic cases and outbreaks of gastroenteritis over several years prior to pandemic emergence (Sdiri-Loulizi et al., 2009; Siebenga et al., 2009) and by publically available putative 'pre-pandemic' sequences, the validity of which we established in the previous chapter (Figure 3.5, Table 3.6). These data raise the questions of where pandemic strains circulate during the years prior to pandemic spread and why we do not regularly observe these strains over the years prior to their emergence. Extensive studies on influenza have demonstrated that epidemic influenza A H3N2 strains originate in South East Asia (Bedford et al., 2015), from where they are spread worldwide via airplane travel (Lemey et al., 2014). In contrast, no consistent source region has been observed for influenza A H1N1 or influenza B, with multiple continents acting as source regions for new viral lineages (Bedford et al., 2015). Previous studies examining norovirus diversity have suggested Africa to harbour a particularly diverse norovirus population, with this high diversity being correlated with the broad range

of HBGAs expressed by different individuals on this continent (Nordgren et al., 2013, 2016). However, this high diversity is defined as the presence of a large range of different genotypes, rather than as a high diversity of different GII.4 variants. GII.4 sequences that do not cluster within any of the main GII.4 strains have been identified in Africa (Mans et al., 2016), despite this region only being sparsely sampled, and this is a characteristic that would be expected of a reservoir region. Additionally, the Hunter 2004 and Sydney 2012 strains were identified in Northern and Southern Africa, respectively, prior to their pandemic emergence (Sdiri-Loulizi et al., 2009; Mans et al., 2016). However, norovirus prevalence is high in all countries worldwide and with the potential for very rapid transmission and links between distant locations via airplane travel and the global food trade, transmission can in theory occur between even distant locations very quickly (de Graaf et al., 2016) and therefore any region could potentially act as a source.

Previous studies have identified individual amino acid changes that alter antigenicity and may have had a role in strain emergence (Lindesmith et al., 2012a; Debbink et al., 2013). For example, the A368E, N373H and G393S substitutions were suggested to be important for the pandemic emergence of Sydney 2012 (Debbink et al., 2013; Allen et al., 2014). However, these studies have typically employed single viruses as representatives of each strain and the conservation of antigenic sites within individual pandemic strains is largely unexplored. Should antigenic sites vary within a pandemic strain, the examination of a single viral sequence cannot result in an accurate picture of strain antigenicity, particularly because different viruses from the same pandemic strain can have different properties (Lindesmith et al., 2008; Debbink et al., 2014). The genetic changes that are essential for the emergence of a pandemic strain are likely those acquired leading to the common ancestor of the strain, as these changes differentiate the viruses within that pandemic cluster from those in other, closely related, clusters that did not cause a pandemic at that point in time. Crucially, our phylogenetic analysis in the previous chapter strongly suggested that the GII.4 strain that emerges to cause a new pandemic is one of a number of low level strains present at that point in time.

Here, we employ phylogeographic methods to trace the spatiotemporal history of the New Orleans 2009 and Sydney 2012 pandemic strains. We demonstrate that each of these strains circulated widely across multiple continents over several years prior to pandemic emergence. There is therefore no evidence for a consistent source region from which

pandemic GII.4 strains emerge and spread. This suggests that environmental factors are very unlikely to drive emergence of new pandemic strains and strengthens our previous suggestion that pandemics are caused by a change in host factors. After the emergence of the pandemic, viruses tend to persist within a continent for long periods of time with a low inter-continental transmission rate, consistent with individuals with symptomatic norovirus infection being unlikely to travel long distances. We show that there is a high level of amino acid diversity within each of the pandemic GII.4 strains and that variable sites often coincide with known epitope sites and/or HBGA-binding regions. It is therefore likely that viruses within a pandemic strain exhibit subtle phenotypic differences. However, viruses within a pandemic strain are likely to exhibit one or more unifying and important phenotypic characteristics. We highlight sites that potentially determine these characteristics by identifying the substitutions that occurred leading to each of the pandemic GII.4 strains. Together, our results suggest extensions to the three stage process of pandemic GII.4 strain emergence proposed in the previous chapter whereby the pre-pandemic strain moves to a new region of antigenic space in the acquisition of changes phase and then circulates widely and acquires subtle antigenic diversity during the diversification phase prior to emerging to cause the new pandemic.

4.3 Materials and Methods

4.3.1 Phylogeographic analyses

We collected datasets containing all of the available capsid sequences on GenBank as of 09/02/2017 for the two most recent pandemic strains, New Orleans 2009 and Sydney 2012. We retained sequences that contained the P2 domain only, as this region is typically highly variable providing phylogenetic signal, but did not retain sequences containing only a small region of the shell domain, as this region is typically conserved and therefore contains little phylogenetic signal. We also removed sequences for which no collection date could be obtained and sequences that were overly divergent based on their sampling date. Sequences overly divergent based on their collection date may contain sequencing errors or have an incorrectly classified date and can therefore mislead tem-

poral evolutionary analyses (Rambaut et al., 2016). There were 565 sequences in the New Orleans 2009 dataset and 708 sequences in the Sydney 2012 dataset. Each dataset exhibited a correlation between root to tip distance and sampling date in a maximum likelihood phylogenetic tree reconstructed using RAxML (Stamatakis, 2014) with the GTR model of nucleotide substitution and gamma rate heterogeneity with four gamma classes. We assessed topological robustness with 1000 bootstrap replicates. Examination of the sampling locations showed that there was typically only a small number of sequences sampled from each country (Table 4.1). We therefore used the continent of collection as the label for each sequence. Labelling each sequence with the continent of collection will enable us to determine the broad location of the virus and the broad patterns of spread while maximising the number of sequences with each label. The Sydney 2012 dataset contains many more sequences from Asia compared with the other continents, while the New Orleans 2009 dataset contains a large number of sequences from Asia and Oceania relative to the other continents (Table 4.1). Should sequences from the same continent typically cluster, continents with more intense sampling are unlikely to artifactually influence estimates of ancestral locations. However, if sequences from different continents are typically interspersed within the tree, an excess of sequences from one continent is likely to artifactually increase support for that continent at ancestral nodes. To assess the clustering of sequences from each continent, we used a maximum likelihood tree for each strain reconstructed with RAxML as above (Figure 4.1). The Sydney 2012 sequences from Asia and the New Orleans 2009 sequences from Asia and Oceania are spread throughout the respective trees, suggesting the more intense sampling in these regions may alter the estimations of ancestral locations. We therefore randomly subsampled 115 of the Sydney 2012 Asia sequences and 95 of each of the New Orleans 2009 Asia and Oceania sequences (with 115 and 95 chosen to match the number of sequences from the next most prevalent region in the respective strain dataset). We carried out three random subsamples and performed all analyses on each of the subsampled datasets. Results were insensitive to the subsampled dataset.

We used the Markov chain Monte Carlo approach implemented in BEAST v2.4.2 (Bouckaert et al., 2014) to reconstruct the phylogeographic history of New Orleans 2009

Country	Number of New Orleans 2009 sequences	Number of Sydney 2012 sequences	Continent
Albania	1	0	Europe
Australia	144	56	Oceania
Bangladesh	8	4	Asia
Brazil	65	0	South America
Canada	6	6	North America
China	21	234	Asia
Denmark	15	23	Europe
France	1	7	Europe
Germany	0	11	Europe
Hungary	1	0	Europe
India	10	3	Asia
Italy	10	4	Europe
Japan	44	87	Asia
Morocco	0	1	Africa
Netherlands	2	2	Europe
New Zealand	0	32	Oceania
Russia	8	2	Europe
Singapore	16	0	Asia
Slovenia	0	56	Europe
South Africa	18	16	Africa
South Korea	29	6	Asia
Spain	5	3	Europe
Sweden	12	0	Europe
Taiwan	16	27	Asia
UK	19	0	Europe
USA	89	109	North America
Vietnam	26	7	Asia

Continent	Number of New Orleans 2009 sequences	Number of Sydney 2012 sequences	
Africa	18	17	
Asia	170	368	
Europe	74	108	
North America	95	115	
Oceania	144	88	
South America	65	0	

Table 4.1: Summary of sequence countries and continents in the New Orleans 2009 and Sydney 2012 datasets. The number of sequences from each country and from each continent are shown within the New Orleans 2009 and Sydney 2012 datasets. The Asia and Oceania sequences within the New Orleans 2009 dataset and the Asia sequences within the Sydney 2012 dataset were subsampled to create the respective phylogeographic datasets.

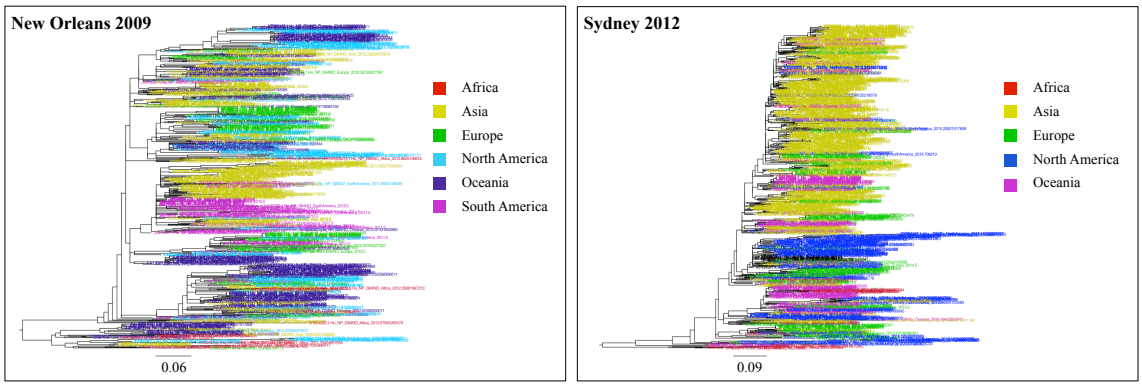


Figure 4.1: Interspersion of seqeunces from each continent in New Orleans 2009 and Sydney 2012. Maximum likelihood trees were reconstructed for the New Orleans 2009 and Sydney 2012 strains. Tips are coloured by the continent on which the sequence was collected. Sequences from different continents are interspersed within each strain. The scale bar represents the expected number of nucleotide substitutions per site.

and Sydney 2012. Sequences were labelled with the most accurate collection date possible; the day of collection if available, the middle of the month of collection if the sampling day was not available and the middle of the year of collection if the month of sampling was not available. We modelled the nucleotide substitution process using the HKY substitution model with amongst site rate variation accomodated with gamma rate heterogeneity with 4 gamma classes. We applied the strict and relaxed lognormal clock models to test for variation in the substitution rate within each clade. In each case, there was strong support to reject a strict clock model in favour of a relaxed lognormal clock model (log10 Bayes factor 66-83) and we therefore used the relaxed lognormal clock model for our inferences. However, the results with the strict clock model are qualitatively very similar, suggesting that potential over-parameterisation due to the large number of branch-specific rate parameters with the relaxed lognormal clock model has not influenced the results. We used a lognormal distribution as the prior on the substitution rate with mean 7×10^{-3} for New Orleans 2009 and 6.4×10^{-3} for Sydney 2012 and standard deviation 0.1 in each case. The mean value was set to the mean posterior estimate of the strain substitution rate from our previous analyses (chapter 3.3.6) and the standard deviation was set to incorporate the 95% HPD of our previous estimate. We used a Bayesian coalescent skyline tree prior for each dataset. We applied a discrete phylogeographic model to describe lineage migrations within each dataset (Lemey et al., 2009). The model includes a symmetric transition matrix for the migration rates, where the rate of migration from location A to location B is the same as that from location B to location A. Relaxation of this constraint will be dis-

cussed below. Each of the parameters within the location transition matrix had a Bayesian stochastic search variable selection (BSSVS) indicator variable, enabling each rate to be either 'on' or 'off'. We set the prior on the number of on rate classes to a Poisson prior with mean 5 for the New Orleans 2009 datasets and mean 4 for the Sydney 2012 datasets, which places 50% of the prior probability on the smallest possible number of location transitions and places high prior probability on a small number of transition rates being on. We assumed an exponential prior with mean 1.0 migration event per lineage per year for the overall rate of geographical transition, and a gamma prior with shape 1.0 and scale 1.0 to each of the relative geographical transition rates in the rate matrix. Three replicate runs were carried out on each dataset with different starting parameters and were run until convergence, as assessed in Tracer v1.6. Runs were combined using LogCombiner v2.2.1 after removal of sufficient burnin and subsampled to obtain between 1500 and 2500 trees which formed the posterior distribution. The MCC tree was identified using TreeAnnotator v2.2.1. Bayesian skyline plots were reconstructed for each strain using Tracer v1.5. We carried out additional runs with New Orleans 2009 without subsampling to compare the length of sample persistence. The models and priors used for these runs were the same as those for the subsampled datasets.

We identified the date of first import into each continent by calculating either the date of the root node if the continent was inferred as the root location or the date of the earliest branch midpoint where the downstream node was inferred to be within the continent. We calculated this date for each continent in each tree in the posterior distribution and calculated the median and 95% HPD of each distribution. We obtain similar results when using the date of the earliest non-root node inferred to be within the continent (i.e. when assuming that the import event occurred at the end of the branch). We used the program posterior analysis of coalescent trees (PACT) v0.9.4 to compute the proportion of viral lineages on each continent through time, the length of persistence of viral lineages within each continent (measured as the average length of time each virus takes to leave its sampled location walking backwards from the tip of the tree), the total and annual migration rates and the proportion of lineages sampled on each continent that were present on each continent going back in time. In each case, estimates were based on the complete posterior distribution of trees.

To identify connected continents, we carried out a Bayes factor test as implemented in SPREAD v1.0.7. Briefly, this test calculates the support for a rate (in our case a symmetric rate) contributing to viral migration by calculating the ratio between the posterior odds of the rate being non zero divided by the prior odds of the rate being non zero. We considered migration rates with Bayes factor support greater than three of being non zero to be well supported (Lemey et al., 2009). We assessed whether New Orleans 2009 and Sydney 2012 have similar global connectivity networks by measuring the Spearman rank correlation coefficient of the migration rates between pairs of continents that are present in each strain dataset.

To determine whether each continent acted as a source or sink of viral lineages, we combined the posterior distribution of trees for each subsampled dataset into a single posterior distribution and counted the total number of import events into the continent and the total number of export events from the continent in each tree in the posterior distribution. We used the ratio of export to import events as a measure of whether each continent donated or received more viral lineages.

4.3.2 Examination of the Sydney 2012 Bayesian skyline plot

To determine whether the two stage increase in the Sydney 2012 Bayesian skyline plot may have been influenced by the addition of the phylogeographic model, we ran BEAST v2.4.2 on the Sydney 2012 subsampled dataset 1 without the inclusion of the phylogeographic model. The other models and priors used were the same as for the phylogeography runs described above, but the sampling locations of each sequence and inference of ancestral locations were not included in the model.

We examined the Bayesian skyline plot at each sampled step in the MCMC chain of the Sydney 2012 subsampled dateset 1 relaxed lognormal clock runs. The two stage increase in relative genetic diversity evident in the Sydney 2012 Bayesian skyline plot could be caused by each sample from the MCMC chain exhbiting this two stage increase or by different sampled steps supporting an increase in relative genetic diversity at different points in time. We therefore divided the MCMC steps into two groups based on whether the first increase in relative genetic diversity (here defined as an increase of more than 100% relative to the population size in the first time slice) occurred before or after

1st January 2012. We reconstructed Bayesian skyline plots for each of these groups using Tracer v1.5. We identified parameters exhibiting a significant difference between the groups using a two sample Kolmogorov Smirnov test.

4.3.3 Identification of potential pandemic enabling substitutions and examination of diversity within GII.4 strains

To identify genetic changes that likely occurred leading to each strain common ancestor, we used the dataset assembled in chapter 2.3.1 containing 2198 GII.4 capsid sequences across all GII.4 strains. We used the nucleotide maximum likelihood phylogenetic tree reconstructed using RAxML (Stamatakis, 2014) and ten bootstrap tree topologies obtained in chapter 2.3.1 for this analysis. Branch lengths within each tree were optimised using the amino acid alignment and the WAG substitution matrix with optimised base frequencies in chapter 2.3.1. In chapter 2.3.1, we also carried out ancestral reconstruction in each tree at the amino acid level using PAML v4.9 (Yang, 2007). We identified the non-synonymous substitutions that occurred leading to each pandemic strain by identifying the substitutions that occurred leading to each strain common ancestor in the maximum likelihood tree and each of the bootstrap tree topologies.

Strain-specific datasets were assembled to contain all of the available capsid sequences for the strain as of 09/02/2017. Viruses were removed if they contained only a small region of the shell domain but retained if they only contained the P2 domain. The sequences from each strain were aligned at the amino acid level using MUSCLE. We reconstructed a nucleotide maximum likelihood phylogenetic tree for each strain using RAxML (Stamatakis, 2014) as above.

For each site inferred to change along the branch leading to a pandemic strain, we calculated the amino acid distribution at this site in that pandemic strain and in the previous pandemic strain. This distribution was calculated using the alignment from the respective strains and so does not take into account the phylogenetic relationships between sequences. The amino acid conservation at each site was also examined using the coloured trees technique, where each tip in the phylogenetic tree is coloured by the amino acid residue in that sequence at the amino acid site of interest. This technique visualises the distribution of amino acids at a single site and enables determination of whether genetic

changes likely occurred on a single occassion, with sequences with the genetic change sharing a common ancestor, or whether there have been multiple independent changes at the site. Homology models were reconstructed for each strain common ancestor and the upstream node using SWISS-MODEL (Bordoli et al., 2008; Biasini et al., 2014). The upstream node was defined as the sequence of the strain common ancestor with each of the substitutions inferred to have occurred leading to the strain common ancestor incorporated. As each of the currently available GII.4 capsid structures consists of only the P domain, we only used the P domain of the reconstructed sequences in the structural modelling. The best fitting template structure was identified using SWISS-MODEL and this single structure was used as the template for homology modelling. Structural analyses were carried out using PyMol v1.74. To visualise residue properties we coloured each residue based on its hydropathy score in the Kyte and Doolittle scale (Kyte and Doolittle, 1982), with more polar residues being shown in blue and more hydrophobic residues being shown in red.

Variable sites were initially identified using the coloured trees technique. Coloured trees were created for each amino acid site in each strain using the capsid strain phylogeny. Variable sites were identified as those sites that exhibited either different amino acid residues in different large clades within the strain or that required multiple nonsynonymous substitutions to explain the distribution of amino acids across the tree at that site. We also used Shannon's entropy to quantify the diversity at each amino acid site in the alignment. The sites identified as highly variable via examination of coloured trees were consistently the sites with the highest Shannon's entropy. As Shannon's entropy does not take the structure of the phylogenetic tree into account, it is possible to obtain comparable diversity estimates from situations requiring vastly different numbers of changes to explain the pattern of residues within the coloured trees. Variable sites, epitope regions (Lindesmith et al., 2012a) and HBGA binding sites (Cao et al., 2007; Singh et al., 2015) were mapped onto the Sydney 2012 P domain structure 4WZT (Singh et al., 2015), which is bound to HBGA type A.

To calculate the diversity at each variable site in New Orleans 2009 and Sydney 2012 through time, we used the MCC tree from the phylogeographic analysis of that strain. We optimised branch lengths within the strain using the amino acid alignment and the WAG substitution matrix with optimised base frequencies using RAxML (Stamatakis, 2014).

We carried out ancestral reconstruction using PAML. The amino acid residue at each site of interest was mapped onto the MCC nucleotide temporally resolved tree as a 'location' label and PACT v0.7.4 was used to reconstruct the proportion of lineages with each amino acid residue through time, as with the geographical location labels described above.

4.4 Results

4.4.1 The two most recent pandemic GII.4 strains circulated widely over several years prior to pandemic emergence

Our results in chapter 3 suggested that each of the pandemic GII.4 strains were present for years prior to causing their respective pandemics. We investigated where the two most recent pandemic GII.4 strains, New Orleans 2009 and Sydney 2012, circulated over the years prior to their pandemic emergence using a phylogeographic analysis (Figure 4.2). The common ancestors of New Orleans 2009 and Sydney 2012 occurred approximately four and seven years prior to pandemic emergence (Figure 4.2), respectively, although as previously discussed it is possible that the true common ancestors of these strains occurred significantly earlier. It is well supported that both strains were imported into each continent prior to the onset of their respective pandemic (Figure 4.3 panels A and B). New Orleans 2009 was likely imported into Africa, Asia and Oceania more than two years prior to the onset of the pandemic, with introduction into North America roughly one year prior to the pandemic onset and into Europe and South America several months prior to pandemic onset. Sydney 2012 was likely imported into Africa, Europe and Oceania more than four years prior to the pandemic onset and into Asia and North America two years or more prior to pandemic onset. These dates, in particular for Sydney 2012, have large confidence intervals on the lower bound of importation due to uncertainty regarding the location of the virus close to the root of tree, likely due to the large time interval between the root and the collection dates of the first viruses to be sampled in each strain. After initial introduction into a continent, it is well supported that New Orleans 2009 and Sydney 2012 continued to circulate within that continent up until the time of pandemic onset and throughout the period in which sequences are available from that continent during the

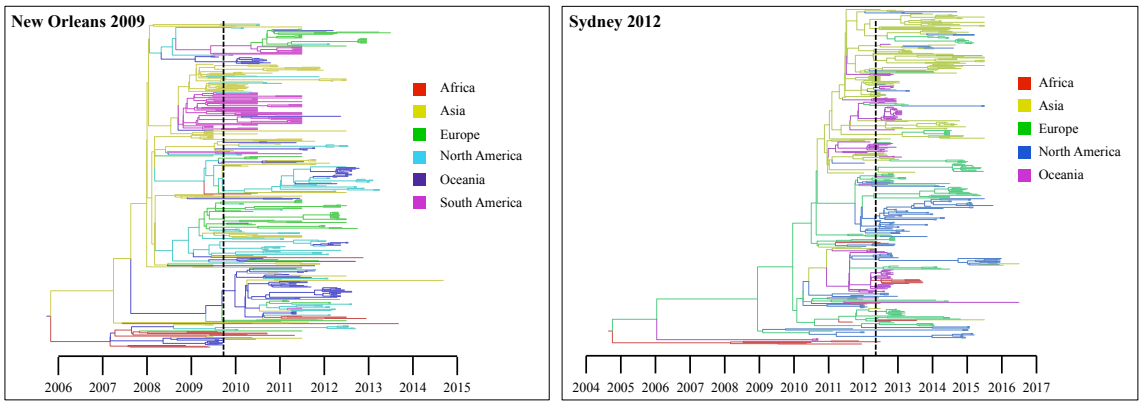


Figure 4.2: Spatiotemporal evolutionary history of New Orleans 2009 and Sydney 2012. Temporal and location resolved MCC trees are shown for the New Orleans 2009 and Sydney 2012 pandemic strains, as reconstructed with BEAST. Branches are coloured by continent. The vertical dashed lines represent the date of the onset of the respective pandemic.

pandemic (Figure 4.3 panels C and D). These results suggest that New Orleans 2009 and Sydney 2012 circulated widely and consistently over several years prior to pandemic onset, despite the respective Bayesian skyline plots suggesting each strain was at low level during this period (Figure 4.3 panels C and D).

The Sydney 2012 Bayesian skyline plot exhibits a two stage increase in relative genetic diversity, with an initial increase in early 2011 followed by another increase in early 2012 coinciding with the onset of the Sydney 2012 pandemic (Figures 4.3, 4.4). This population history may indicate that Sydney 2012 emerged in a two stage process. This two stage increase is not due to the inclusion of the phylogeographic model, as we observe the same results with the same dataset without including the phylogeographic model (Figure 4.4). Indeed, this two stage increase was also evident in our previous Sydney 2012 Bayesian skyline plots (Figure 3.3), although the phylogeographic dataset exhibits a greater increase in relative genetic diversity in the first stage compared with our previous analysis. We examined the skyline of each sampled step in the MCMC chain and identified two major patterns in the relative genetic diversity through time. Most sampled steps (roughly 75%) exhibit a two stage increase in relative genetic diversity, as is the case in the Bayesian skyline plot calculated on all MCMC chain samples (Figure 4.4). We therefore defined this pattern as scenario 1. A smaller number of sampled steps exhibit a single increase in relative genetic diversity in 2012, coinciding with the time of the second increase in scenario 1 (Figure 4.4). We defined this pattern as scenario 2. The

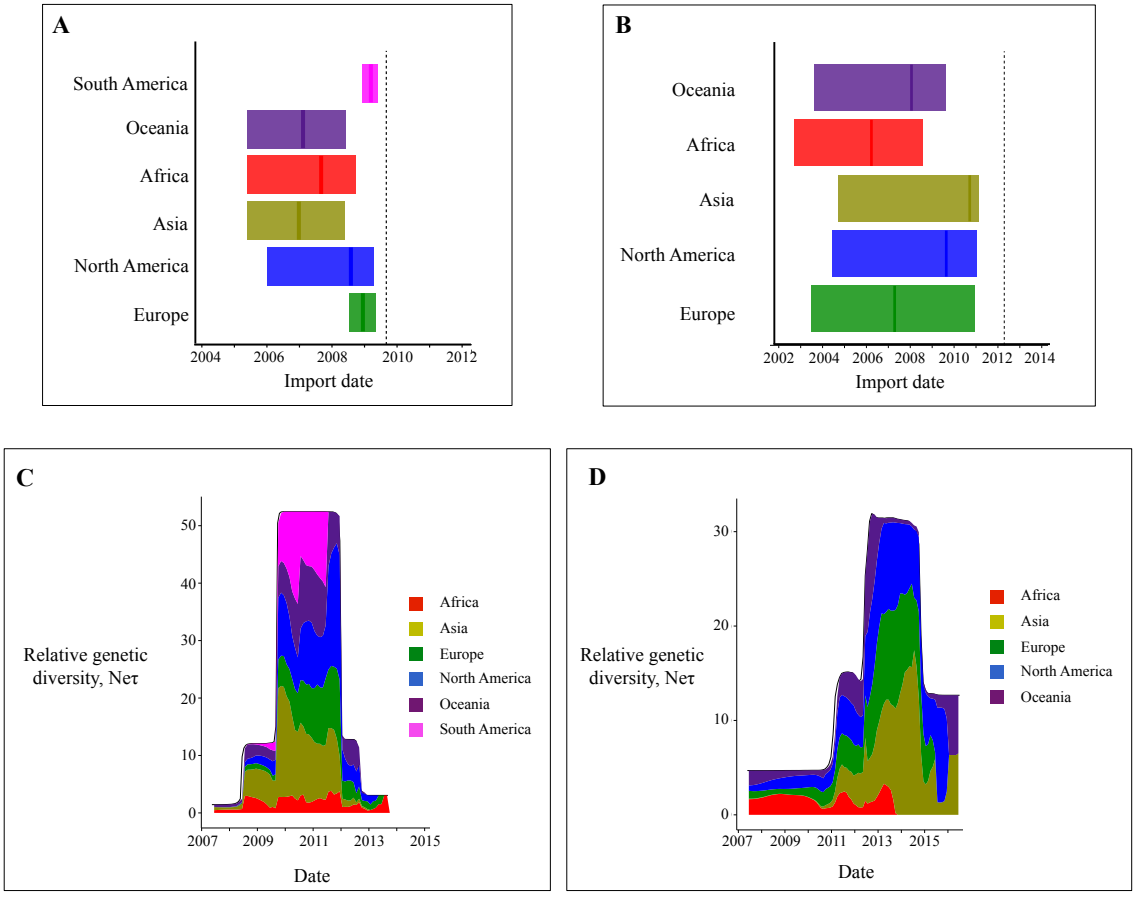


Figure 4.3: New Orleans 2009 and Sydney 2012 circulated widely and consistently over several years prior to pandemic emergence. (**A** and **B**) Summary of continent importation dates for New Orleans 2009 and Sydney 2012. The distribution of the first date of importation was calculated for each continent for New Orleans 2009 (**A**) and Sydney 2012 (**B**). The vertical line represents the mean date and the shaded area the 95% HPD. (**C** and **D**) Summary of the spatial distribution of New Orleans 2009 and Sydney 2012 lineages through time. The proportion of viral lineages on each continent is plotted through time for New Orleans 2009 (**C**) and Sydney 2012 (**D**) and is scaled to the relative genetic diversity (Neτ) from the Bayesian skyline plot for each strain. The locations through time are plotted as a stacked area plot.

Bayesian skyline plot consists of a number of time slices (in this case five) with a different scaled effective population size in each time slice. During the MCMC chain, the length of each time slice and the scaled effective population size in each time slice are integrated over in the form of bGroupSizes and bPopSizes parameters, respectively. There is a significant difference ($p < 0.05$) between scenario 1 and scenario 2 in the posterior distributions of the four earliest bGroupSizes parameters and between the three earliest bPopSizes parameters. Therefore there is a significant difference between the scenarios in the length of time of the four earliest time slices in the Bayesian skyline plot and between the scaled effective population size within the three earliest time slices. In particular, the

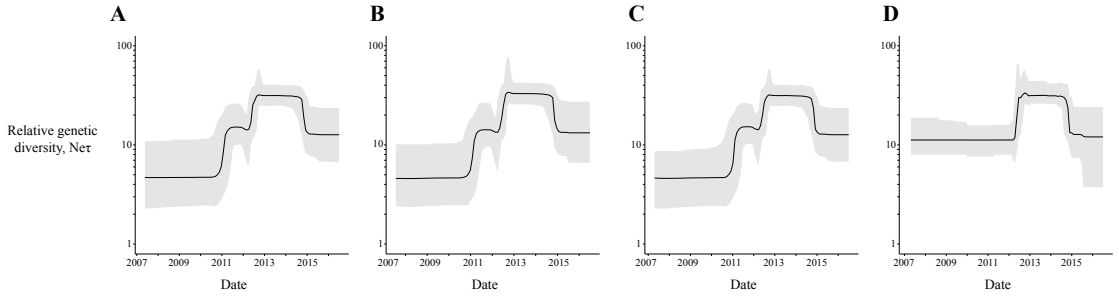


Figure 4.4: The population history of Sydney 2012. Bayesian skyline plots reconstructed for the Sydney 2012 pandemic strain. (**A**) Bayesian skyline plot of the subsampled Sydney 2012 dataset with phylogeography. (**B**) Bayesian skyline plot of the subsampled Sydney 2012 dataset without phylogeography. (**C**) Bayesian skyline plot of MCMC chain samples that exhibit the first increase in relative genetic diversity pre-2012, defined here as scenario 1. Here, we defined an increase in relative genetic diversity as an increase in Neτ of more than 100% relative to the baseline value. (**D**) Bayesian skyline plot of MCMC chain samples that exhibit an increase in relative genetic diversity in or after 2012, defined here as scenario 2.

earliest bGroupSizes parameter is significantly larger in scenario 2, meaning there are more coalescent events within the first time slice and therefore delaying the time at which the second time slice begins. As the population size increases in the second time slice, this delays the time of the increase in population size. In contrast, the second time slice occurs earlier in scenario 1, with this time slice having an intermediate population size between that in time slices one and three, resulting in the two stage increase in relative genetic diversity. This second time slice in scenario 1 has a very similar scaled effective population size to that in the first time slice in scenario 2 (Figure 4.4). There is, however, no significant difference between the distributions of posterior probability or likelihood of scenario 1 and scenario 2. Therefore there is no siginificant difference in how well these two scenarios fit the data and, despite more MCMC steps supporting scenario 1, there is not strong support in favour of a two stage increase in relative genetic diversity (one pre-pandemic and one at the onset of the pandemic) over a single stage increase in relative genetic diversity at the onset of the pandemic.

4.4.2 Phylogeography supports rare intercontinental transmission with long intracontinental persistence

Both maximum likelihood (Figure 4.1) and Bayesian phylogenies (Figure 4.2) demonstrate that sequences sampled from the same geographical region do not always clus-

ter together. There are sequences sampled from each continent spread throughout the New Orleans 2009 and Sydney 2012 phylogenies, indicating historical intercontinental transmission events. Indeed, each continent contains a high diversity of lineages from each pandemic strain, with viruses from multiple clades within a strain often being found within a continent in a single season (Figure 4.2). However, viral lineages typically persist within a continent for a long period of time, with the average persistence of New Orleans 2009 and Sydney 2012 on a continent being more than two years (Figure 4.5). The average persistence of New Orleans 2009 lineages is similar for each continent, although this strain typically persisted for longer within Asia compared with the other continents. The average persistence of Sydney 2012 lineages is also similar for most continents, with evidence of shorter persistence in Oceania. However, the majority of Sydney 2012 sequences in our dataset sampled in Oceania were collected during the 2012-2013 season, with very few sequences from Oceania in 2014, which is likely to have decreased persistence in this continent. In total, 86% and 85% of New Orleans 2009 and Sydney 2012 lineages, respectively, were imported into their continent of sampling more than one year prior to sampling (Figure S4.1). If viruses from different continents are interspersed within the phylogenetic tree, subsampling these viruses may inflate persistence estimates due to reducing the number of migration events. However, the length of persistence is very similar within each subsampled New Orleans 2009 dataset and in a New Orleans 2009 dataset without subsampling (Figure S4.2), indicating that subsampling is unlikely to have greatly influenced estimates of sample persistence. There is no correlation between the average persistence in each continent between New Orleans 2009 and Sydney 2012 (Figure S4.3).

The intercontinental transmission rate is similar for New Orleans 2009 (0.2946 migrations/lineage/year) and Sydney 2012 (0.2561 migrations/lineage/year). This rate corresponds to roughly one in four lineages migrating to a different continent each year, suggesting that intercontinental transmission is not a very frequent event. Estimates of the annual migration rate indicate that the highest rate of viral migration occurs in the year of and the years preceding pandemic spread, with the migration rate decreasing during the pandemic period (Figure 4.6).

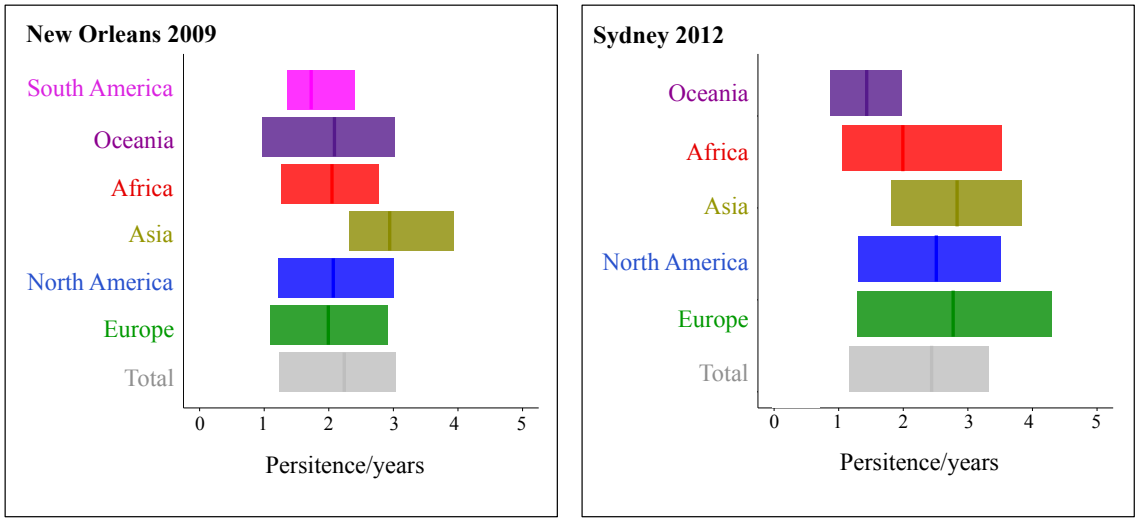


Figure 4.5: Summary of lineage persistence within each continent. For each tip virus, we calculated the length of time it took to leave the sampling continent walking backwards up the phylogenetic tree. The continental persistence was calculated by combining the length of persistence of each tip virus from that continent in each tree in the posterior distribution. The mean of the distribution is shown by the vertical line and the 95% HPD is shown by the shaded area. The total persistence is the average persistence across all continents.

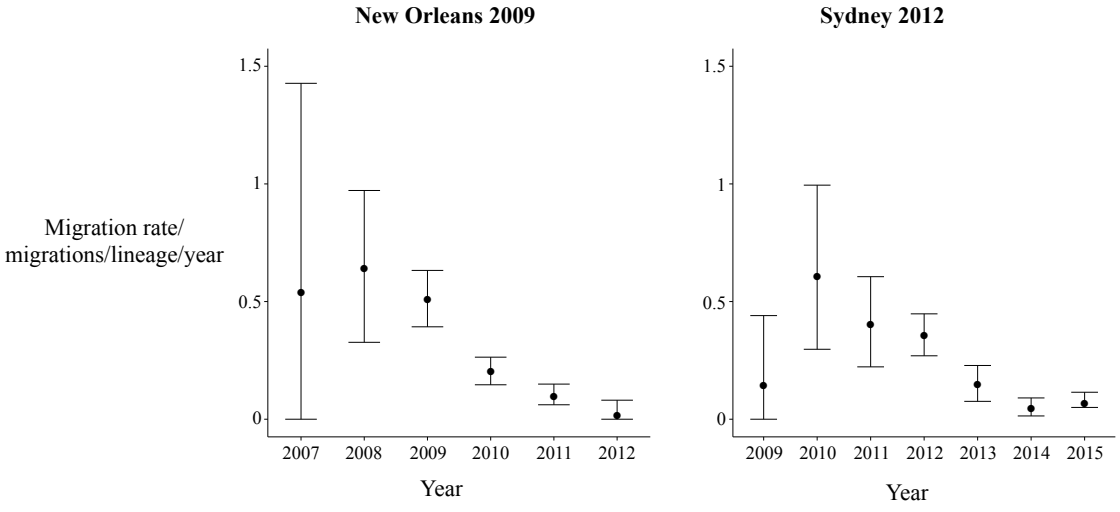


Figure 4.6: Annual migration rates for New Orleans 2009 and Sydney 2012. The annual migration rate was calculated as the rate at which the location label changes within each year. The point represents the mean migration rate for the year and the error bars represent the 95% HPD.

4.4.3 Identification of frequent inter-continental transmission pathways and Asia as a source of viral lineages

We next investigated the global connectivity network within the New Orleans 2009 and Sydney 2012 strains (Figure 4.7). Continents are generally well-connected, with each continent having at least one well supported connection in each dataset and most continents having more than one well supported connection (Figure 4.7). Therefore while the inter-continental transmission rate is low overall, there is evidence for occasional migration events between a large number of continents. We do not find evidence for New Orleans 2009 and Sydney 2012 sharing the same global connectivity network (Spearman rank correlation coefficient 0.4704), although this estimate may be biased by the New Orleans 2009 dataset additionally containing sequences from South America. While several continent pairs exhibit greatly different migration rates when comparing New Orleans 2009 and Sydney 2012, the migration rate is high between Europe and North America and between Asia and Oceania in both strains (Figure 4.8).

We next identified whether each continent acted as a source (i.e. exports more viral lineages than it imports) or a sink of viral lineages or whether the number of import and export events is roughly equal (Figure 4.9). While most continents export and import a similar number of viral lineages or act as lineage sinks, Asia acted as a source of viral lineages in both New Orleans 2009 and Sydney 2012 (Figure 4.9).

4.4.4 Identification of potential pandemic-enabling substitutions

The phylogenetic trees reconstructed as part of the phylogeographic analysis exhibit the same pattern as those reconstructed in chapter 3 where New Orleans 2009 and Sydney 2012 diversify into a large number of lineages prior to pandemic emergence. In chapter 3 we demonstrated that this is true of the five most recent pandemic strains and introduced a model of strain emergence where the viral genetic changes that enable the pandemic to occur are acquired years prior to pandemic onset (Figure 3.6, stage 1).

We next aimed to identify substitutions that may have been important for the pandemic emergence of each strain by carrying out a thorough examination of amino acid change across the GII.4 tree. This work was carried out as part of a collaboration and aimed to identify substitutions that could be tested experimentally to determine their ef-

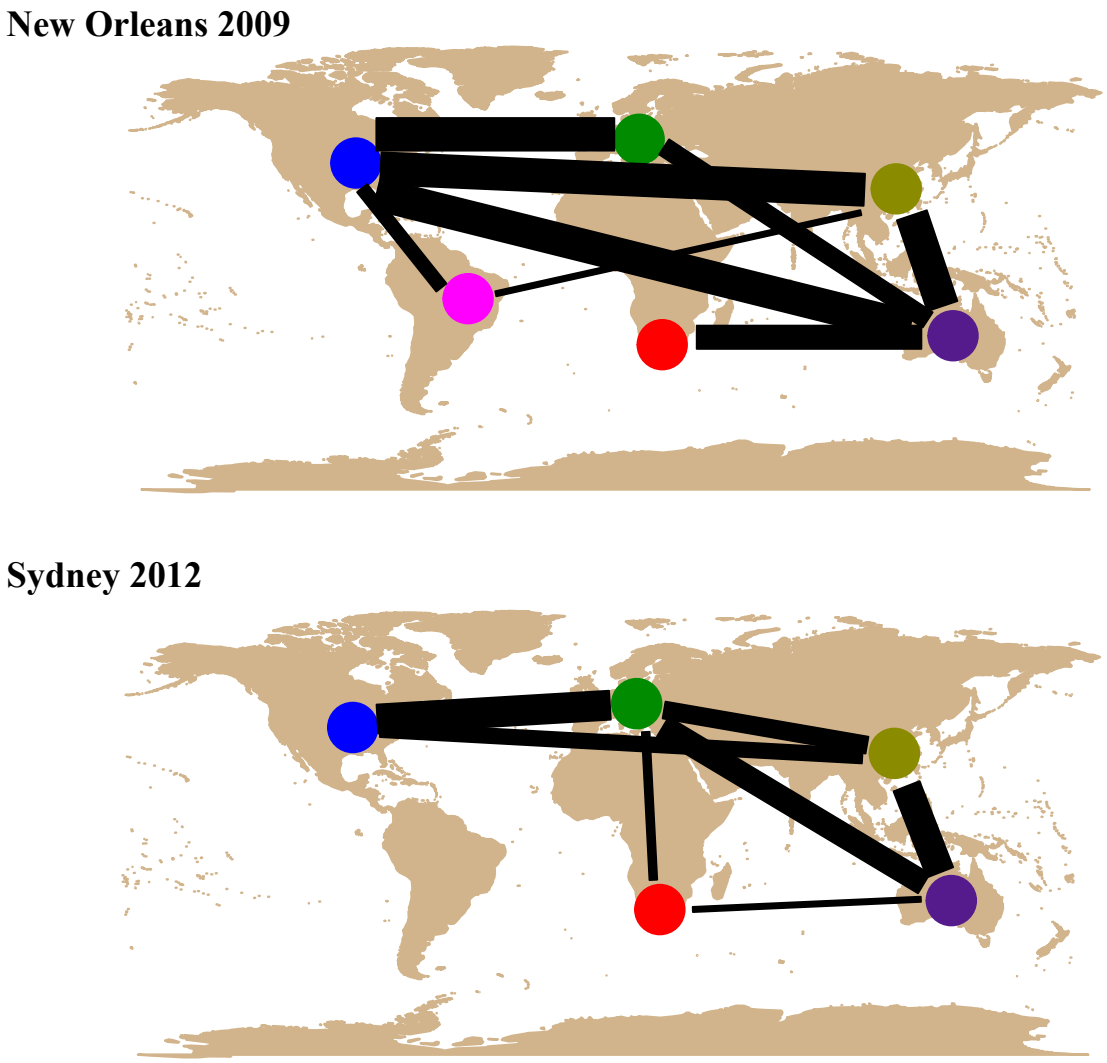


Figure 4.7: The global connectivity network for New Orleans 2009 and Sydney 2012. Lines are shown between continents with support (Bayes factor > 3) for connection in at least two of the three subsampled datasets. The thickness of the line is proportional to the median estimate of the migration rate between the pair of continents.

fect on viral antigenicity. We would expect substitutions important for pandemic emergence to occur along the branch in the phylogenetic tree leading to the pandemic clade and result in a different amino acid residue(s) to that in the previous pandemic strain. In chapter 3, we demonstrated that there are multiple low level GII.4 lineages present through time and that multiple low level GII.4 strains were present at the emergence of each pandemic strain, for example at the onset of the Hunter 2004 pandemic, the lineages leading to Den Haag 2006, Yerseke 2006, Apeldoorn 2007, New Orleans 2009 and Sydney 2012 were likely all present but at low level (Figure 3.1, Table 3.4). The substitutions that occurred leading to the pandemic clade are specific to that pandemic cluster and are

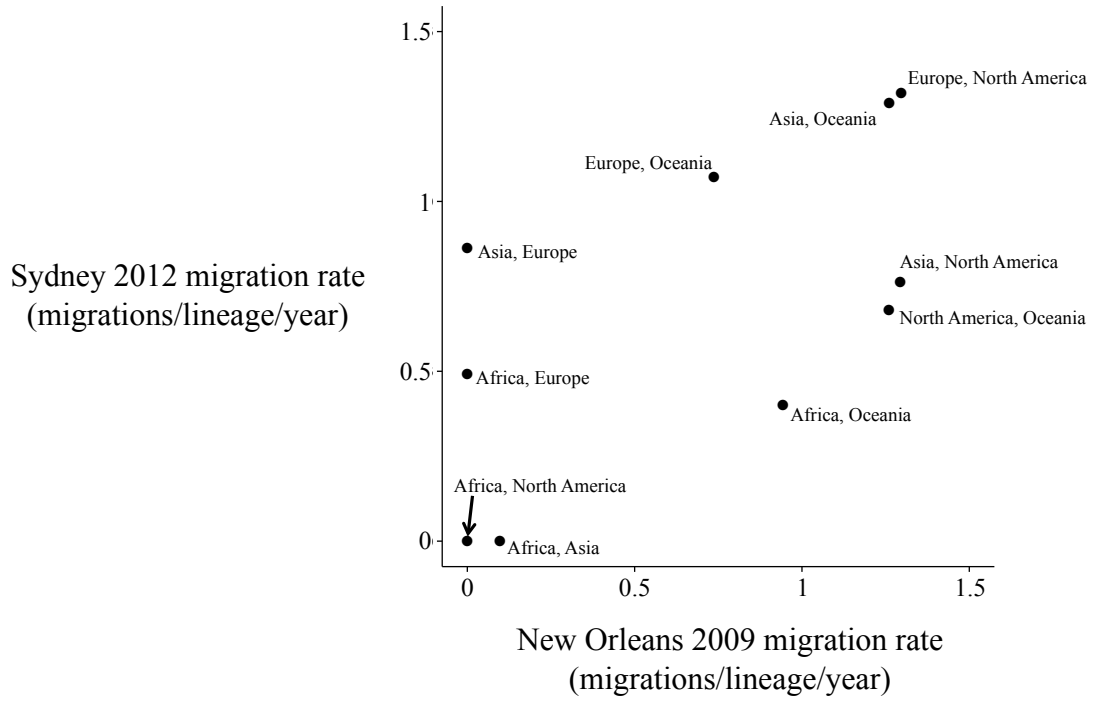


Figure 4.8: Comparison of inter-continental migration rates in New Orleans 2009 and Sydney 2012. The median migration rate is plotted for each pair of continents in New Orleans 2009 and Sydney 2012. Only continent pairs that are present in both datasets are included.

therefore likely to have enabled the emergence of that cluster over the other lineages that were also present at the time. We therefore identified the nonsynoymous substitutions that occurred leading to the common ancestor of each of the pandemic clusters and compared the distribution of residues at each of these sites to that in the preceding pandemic strain. As the topology of the GII.4 tree is poorly supported in several regions, we carried out our analyses on the maximum likelihood tree and ten bootstrap tree topologies from the capsid dataset assembled in chapter 2.3.1.

Sydney 2012

The topology of the Sydney 2012 clade has been consistent in each of our analyses with this strain, with the common ancestor predicted to occur in late 2004 (95% HPD late 2001-mid 2007). However, there are two clades of sequences collected in South Africa and New Zealand that branch from or close to the root of the clade (Figure 4.10). The sequences within these clades were mostly collected in 2010 and 2011 and there is therefore no evidence that these viruses contributed to the Sydney 2012 pandemic. All of the viruses

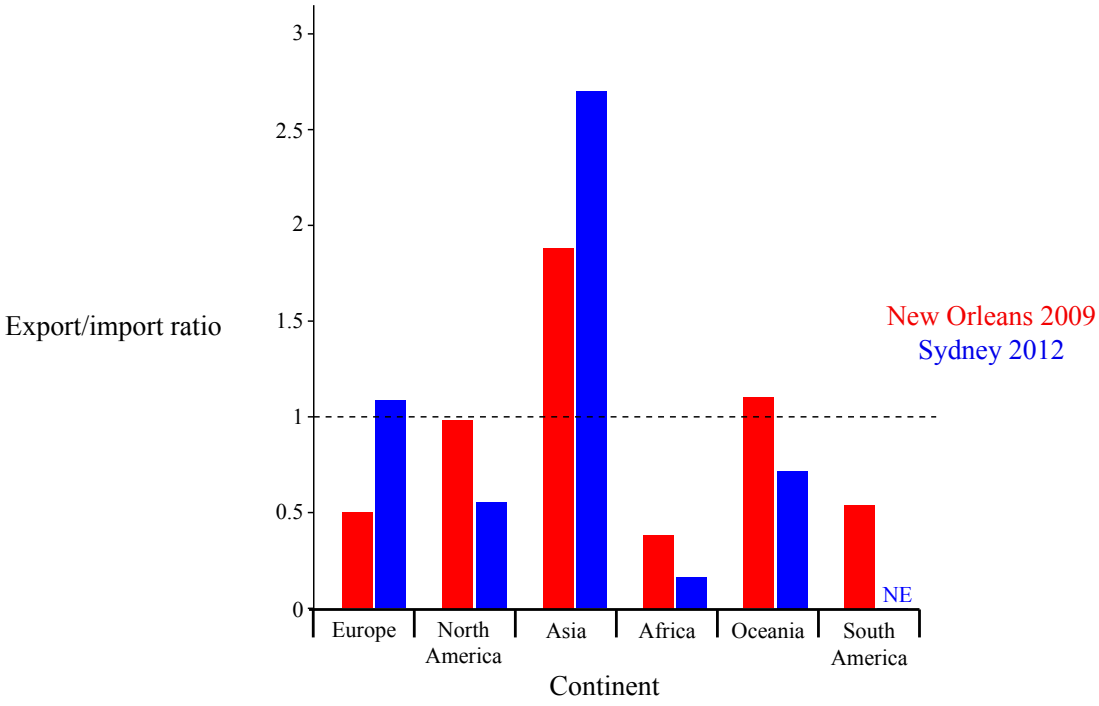


Figure 4.9: Asia acted as a source of New Orleans 2009 and Sydney 2012 lineages. We counted the total number of viral lineage import and export events for each continent and calculated the export/import ratio. This ratio is plotted for each continent for New Orleans 2009 (red) and Sydney 2012 (blue). The horizontal dashed line at the export/import ratio of 1 shows an equal number of export and import events. Therefore continents with ratios under this line are sinks while continents with ratios above this line are sources. NE - not estimated due to sequences from that continent not being included in the dataset.

that contributed to the Sydney 2012 pandemic form a monophyletic cluster that coalesces to a common ancestor in late 2008 (95% HPD mid 2007-late 2009) (Figure 4.10). We therefore defined two nodes within the Sydney 2012 clade: node 1 which is the common ancestor of all Sydney 2012 viruses and node 2 which is the common ancestor of the pandemic Sydney 2012 viruses (Figure 4.10). It is possible that pandemic-enabling substitutions were acquired leading to node 1 and the South Africa and New Zealand clades became extinct due to stochastic factors. However, it is also possible that substitutions important for the pandemic emergence of Sydney 2012 were acquired leading to node 2 and the South Africa and New Zealand clades did not contribute to the Sydney 2012 pandemic because they did not have one or more of the important substitutions. Each of the clades downstream of node 2 persisted throughout the pandemic, with no evidence of one or more lineages outcompeting other lineages, suggesting that important substitutions had been acquired by node 2 (Figure 4.10).

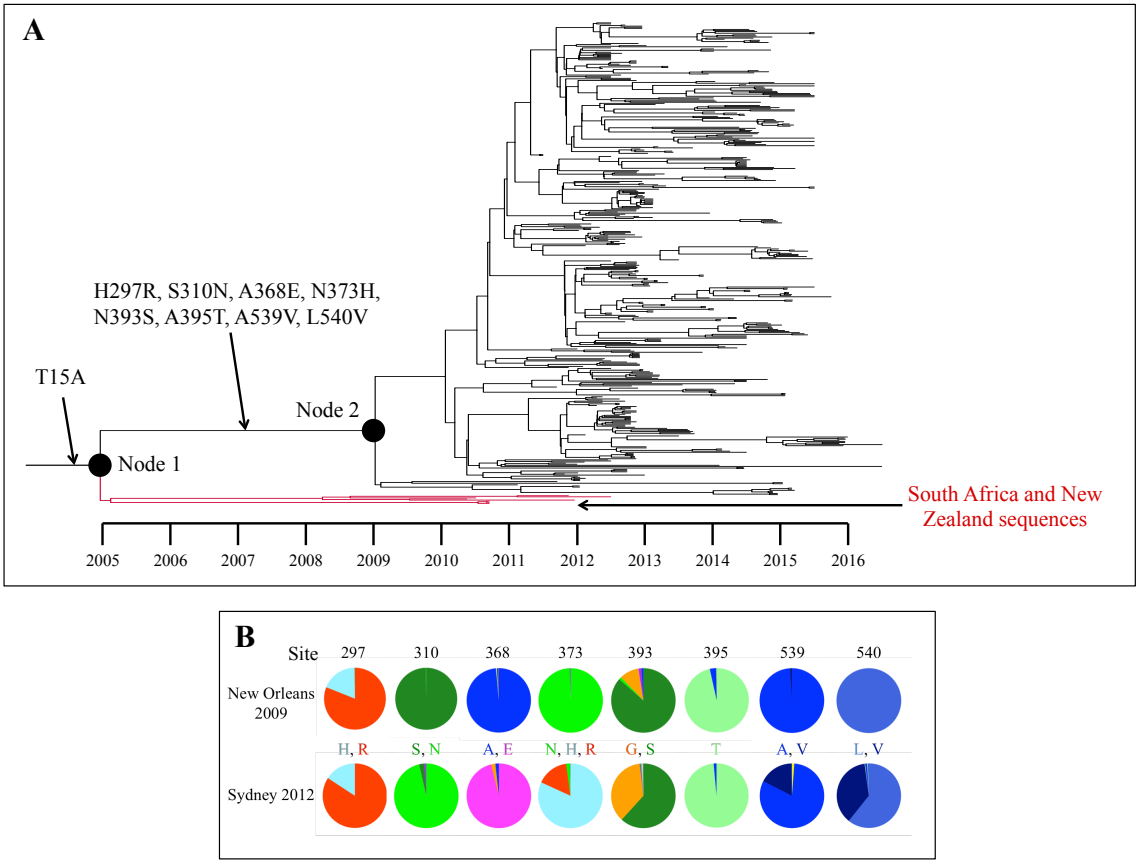


Figure 4.10: Nonsynonymous substitutions leading to the Sydney 2012 clade. (**A**) MCC tree of the Sydney 2012 strain, as in figure 4.2 but without the branches labelled by location. We identified node 1 as the common ancestor of all Sydney 2012 viruses and node 2 as the common ancestor of the pandemic GII.4 viruses. All of the viruses that contributed to the Sydney 2012 pandemic are downstream of node 2. The South Africa and New Zealand clades that diverge from or close to the root of the tree are labelled in red. The nonsynonymous substitutions that occurred leading to node 1 and node 2 are labelled on the respective branch. (**B**) Comparison of the amino acid distribution at each of the sites that change leading to node 2 in New Orleans 2009 and Sydney 2012. Amino acid residues at each site are shown between the distributions and residues with similar properties are shown in similar colours. The same residues are found in New Orleans 2009 and Sydney 2012 at sites 297, 393, 395, 539 and 540.

We infer that a single nonsynonymous substitution occurred along the branch leading to node 1: T15A (Figure 4.10). Site 15 is located in the shell domain and has been identified as evolving under positive selection in GII.4 (Bok et al., 2009), although the functional importance of substitutions at this site is unknown. As this site is located within the shell domain, it is very unlikely that substitutions at this site would alter viral antigenicity.

We infer that eight nonsynonymous substitutions occurred along the branch leading to node 2 (Figure 4.10). However, the amino acid residues at sites 297, 393, 395, 539 and 540 in Sydney 2012 are the same as those in the preceding pandemic strain, New Orleans 2009 (Figure 4.10). This suggests that substitutions at these sites were not the key drivers of pandemic emergence of Sydney 2012, but rather the key substitution was one or more of S310N, A368E and N373H (Figure 4.11). Sites 368 and 373 are located within blockade epitope A (Lindesmith et al., 2012a) and site 310 has previously been demonstrated to alter viral antigenicity, potentially through regulating antibody access to an epitope (Lindesmith et al., 2014). Sites 310 and 368 are largely conserved as asparagine and glutamic acid, respectively, within the Sydney 2012 pandemic clade, while site 373 is largely conserved as histidine but with arginine also frequently present at this site in the pandemic clade (Figures 4.10, 4.11). However, both residues present at site 373 in Sydney 2012 are different to the residue at this site in New Orleans 2009 (Figure 4.10). The substitutions at sites 368 and 373 have previously been suggested to be important for the emergence of the Sydney 2012 pandemic (Debbink et al., 2013; Allen et al., 2014) and homology modelling suggests that the substitutions at these sites resulted in a change in the structure of blockade epitope A (Figure 4.11). We therefore hypothesise that substitutions within blockade epitope A may have contributed to the pandemic emergence of Sydney 2012 and identify the substitutions S310N, A368E and N373H for experimental validation of their effect on viral antigenicity.

New Orleans 2009

There were three nonsynonymous substitutions that occurred leading to the common ancestor of the New Orleans 2009 pandemic strain in the majority of tested tree topologies: T294A, A340T and A359S (Figure 4.12). Each of these sites has a different distribution of amino acid residues in New Orleans 2009 compared with that in the previ-

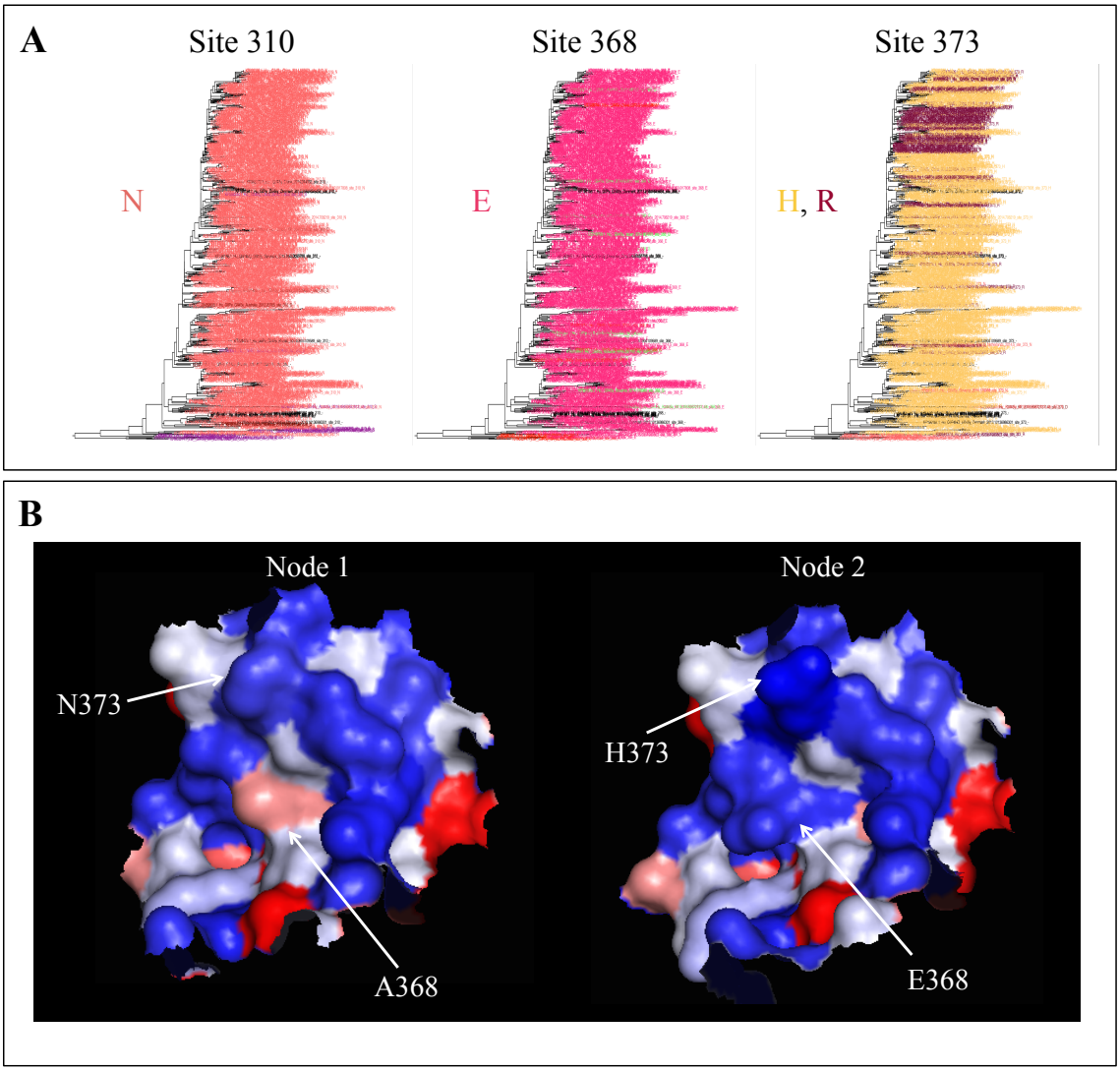


Figure 4.11: Sites that may have enabled the pandemic emergence of Sydney 2012. (**A**) Conservation of sites 310, 368 and 373 within the Sydney 2012 strain. Each tip within a maximum likelihood tree of 782 Sydney 2012 capsid sequences is coloured by the residue at the respective site. The common amino acid residues present at each site are labelled next to the respective tree. The site was not sequenced in tips coloured black. Sites 310 and 368 are largely conserved, while site 373 is mostly histidine with arginine also present at this site. (**B**) The change within epitope A between Sydney 2012 node 1 and node 2. Residues within 12 *Å* of site 368 are shown, with sites in this region forming blockade epitope A (Lindesmith et al., 2012a). Residues are coloured by hydrophobicity on the Kyte and Doolittle scale (Kyte and Doolittle, 1982), with blue residues being more polar and red residues more hydrophobic. Sites 368 and 373 that are located within blockade epitope A and change between node 1 and node 2 are highlighted. The substitution at site 368 results in a change in residue size and polarity, while the substitution at site 373 results in a change in residue size.

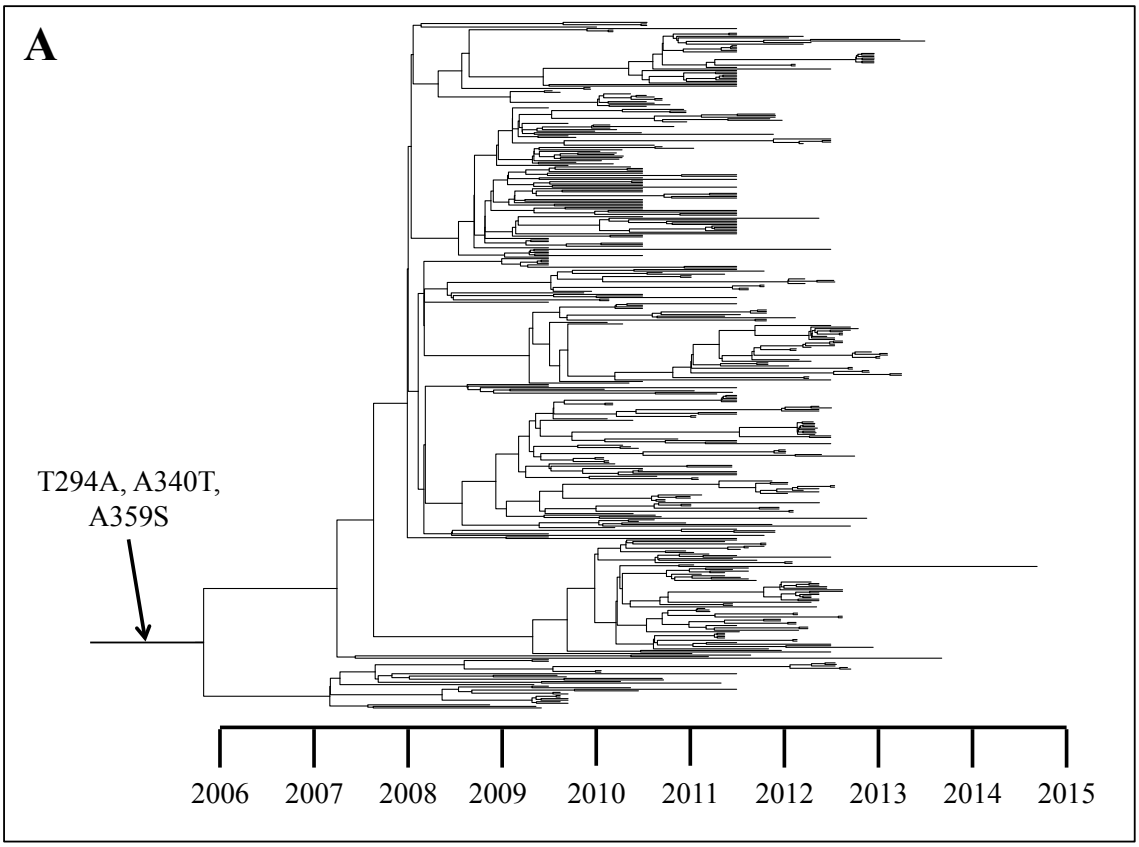
ous pandemic strain, Den Haag 2006 (Figure 4.12). Sites 294 and 340 are located within blockade epitope A and putative epitope C (Lindesmith et al., 2012a), respectively.

While site 294 changes leading to the New Orleans 2009 common ancestor, it is highly variable within the New Orleans 2009 clade, with multiple substitutions between adenine, proline and serine (Figure 4.13). Interestingly, site 294 was largely conserved as adenine in Den Haag 2006 (Figure 4.12) and New Orleans 2009 lineages with adenine at this site appeared to circulate at lower prevalence during the pandemic than viruses with proline or serine at this site (Figure 4.13). If site 294 was vital to enable the pandemic emergence of New Orleans 2009, we would not expect to see New Orleans 2009 viruses with adenine at this site. The lower prevalence of lineages with adenine at site 294 suggests that this site may have had a more minor role in, or very little influence on, strain emergence.

Site 340 is largely conserved as threonine in the New Orleans 2009 clade in contrast to the glycine residue dominant at this site in Den Haag 2006 (Figures 4.12, 4.13). However, a convergent adenine to threonine substitution occurred early in the pandemic Sydney 2012 clade and Sydney 2012 viruses with threonine at site 340 were likely present at the onset of the New Orleans 2009 pandemic. This suggests that either site 340 did not have a major influence on the pandemic emergence of New Orleans 2009, or that Sydney 2012 had acquired additional changes that decreased the advantage provided by the substitution at site 340. Site 359 is conserved as serine in New Orleans 2009 which differs to the adenine and threonine found at this site in other strains. While site 359 is not within a previously characterised epitope, it is close to epitope A (Figure 4.13). Therefore, while there is no clear candidate substitution for driving the New Orleans 2009 pandemic, we identify sites 294, 340 and 359 for experimental investigation.

Den Haag 2006

The clustering of Den Haag 2006 within the GII.4 clade is uncertain. In particular, the clustering of Farmington Hills 2002, Lanzhou 2002, Asia 2003, Den Haag 2006, the Hunter lineage (leading to Hunter 2004 and Yerseke 2006) and the Apeldoorn lineage (leading to Apeldoorn 2007, New Orleans 2009 and Sydney 2012) is typically poorly supported (Figure 2.1). This poorly supported topology is likely partly due to convergent amino acid substitutions and correspondingly several substitutions occurred leading to the



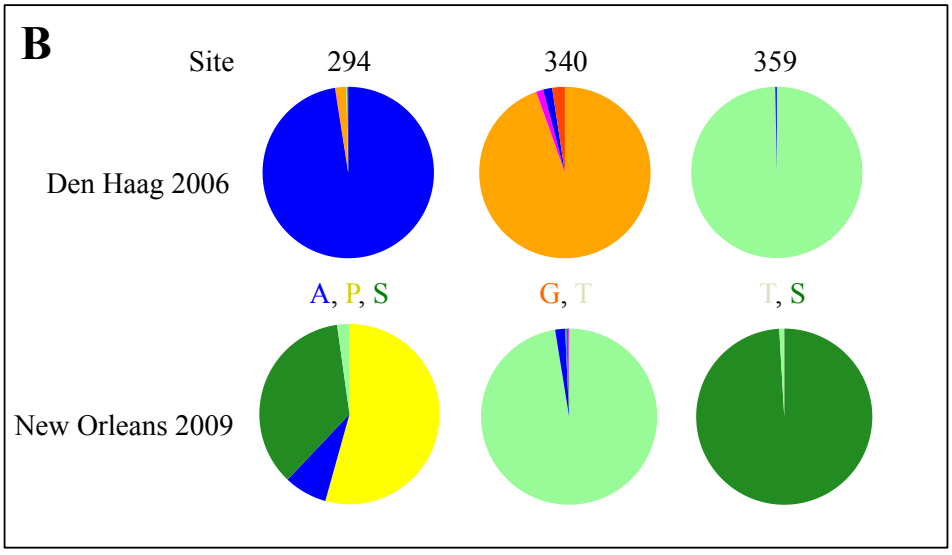


Figure 4.12: Nonsynonymous substitutions leading to the New Orleans 2009 clade. (**A**) MCC tree of the New Orleans 2009 strain, as in figure 4.2 but without the branches labelled by location. The nonsynonymous substitutions that occurred leading to the common ancestor of the New Orleans 2009 strain are labelled. (**B**) Comparison of the amino acid distribution at each of the sites that changed leading to the common ancestor of New Orleans 2009 with that in the previous pandemic strain, Den Haag 2006. Amino acid residues at each site are shown between the distributions and residues with similar properties are shown in similar colours.

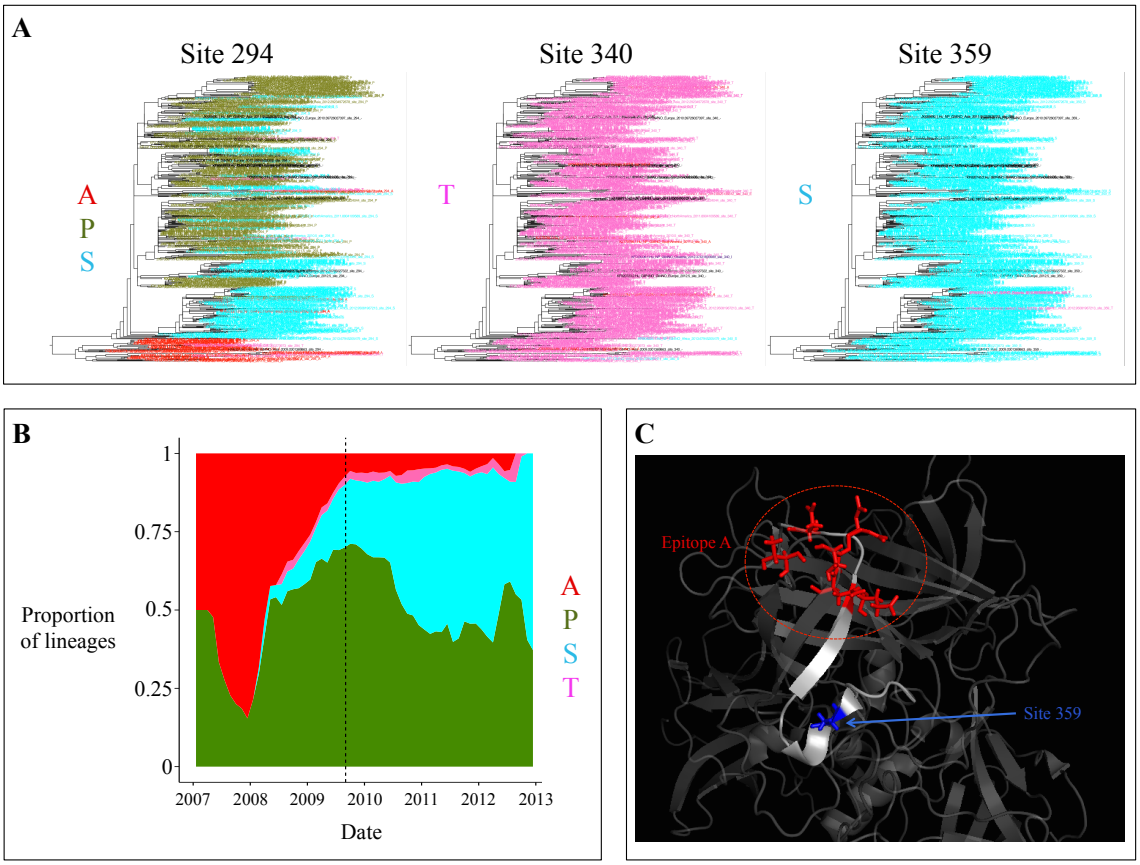


Figure 4.13: Sites that may have enabled the pandemic emergence of New Orleans 2009. (**A**) Conservation of sites 294, 340 and 359 within the New Orleans 2009 strain. Each tip within a maximum likelihood tree of 540 New Orleans 2009 capsid sequences is coloured by the residue at the respective site. The common amino acid residues present at each site are labelled next to the respective tree. Tips coloured black indicate that the site was not sequenced in that sample. Sites 340 and 359 are largely conserved while site 294 is variable. (**B**) The proportion of New Orleans 2009 lineages with each residue at site 294 through time. The proportion of lineages with A, P, S and T at site 294 is plotted through time. This was calculated from ancestral reconstruction on a temporally-resolved phylogenetic tree. Therefore all lineages and not only tip viruses are considered. The proportions are shown as a stacked area plot. (**C**) The location of site 359 (blue) relative to epitope A (red). The capsid structure linking site 359 to residues in epitope A is shown in grey, while the rest of the capsid structure is partially transparent.

common ancestor of Den Haag 2006 in some but not all bootstrap tree topologies. Due to the poorly supported tree topology, we examined the distribution of amino acids across the GII.4 clade at each site that was inferred to change leading to the Den Haag 2006 common ancestor in one or more of the 11 tested tree topologies. There is strong support that the substitutions T15A, Q306L, H357P, N372E and G378H occurred leading to the Den Haag 2006 common ancestor, with the residues at sites 306, 357, 372 and 378 in Den Haag 2006 being unique within the GII.4 clade (Figure 4.14). Additionally, the substitutions P174S, H297R, R339K, S352Y, V356A, N407S and G413V were inferred to occur leading to the Den Haag 2006 common ancestor in more than half of the tested tree topologies (Figure 4.14). Each of these sites has a convergent substitution in one or more other strains, likely explaining why they only occurred leading to the Den Haag 2006 common ancestor in a subset of the tested topologies. There are additional substitutions that occurred leading to the Den Haag 2006 common ancestor in a small number of tested tree topologies, with these sites exhibiting the same residue in Den Haag 2006 as in several of the other strains.

We further considered the 12 sites that changed leading to the Den Haag 2006 common ancestor in more than half of the tested tree topologies. Each of these sites exhibits a different distribution of residues in Den Haag 2006 compared with the previous pandemic strain, Hunter 2004 (Figure 4.14). With the exception of site 357, each of these sites is largely conserved in the Den Haag 2006 clade (Figure 4.14). Sites 15 and 174 are located in the shell domain and are therefore unlikely to alter the antigenicity of the virus. However, sites 297, 339, 352, 356, 357, 372, 378, 407 and 413 are all surface exposed (Figure 4.15). Sites 297 and 372 are in blockade epitope A, while sites 407 and 413 are in blockade epitope E (Lindesmith et al., 2012a) and sites 352, 356 and 357 are located close to both blockade epitopes A and E. Homology models of the common ancestor of the Den Haag 2006 strain and the upstream ancestor suggest that these substitutions may have altered the structure of blockade epitope A, blockade epitope E and the region in between these epitopes (Figure 4.16). Additionally, while site 306 is not surface exposed, it is located close to site 310 which can regulate antibody access an as yet unidentified epitope (Lindesmith et al., 2014). While site 306 is conserved as leucine within Den Haag 2006, this site is completely conserved as glutamine in all other GII.4 strains.

Our results therefore suggest that there was a large change in surface structure leading to Den Haag 2006, with the potential for changes in at least two blockade epitope re-

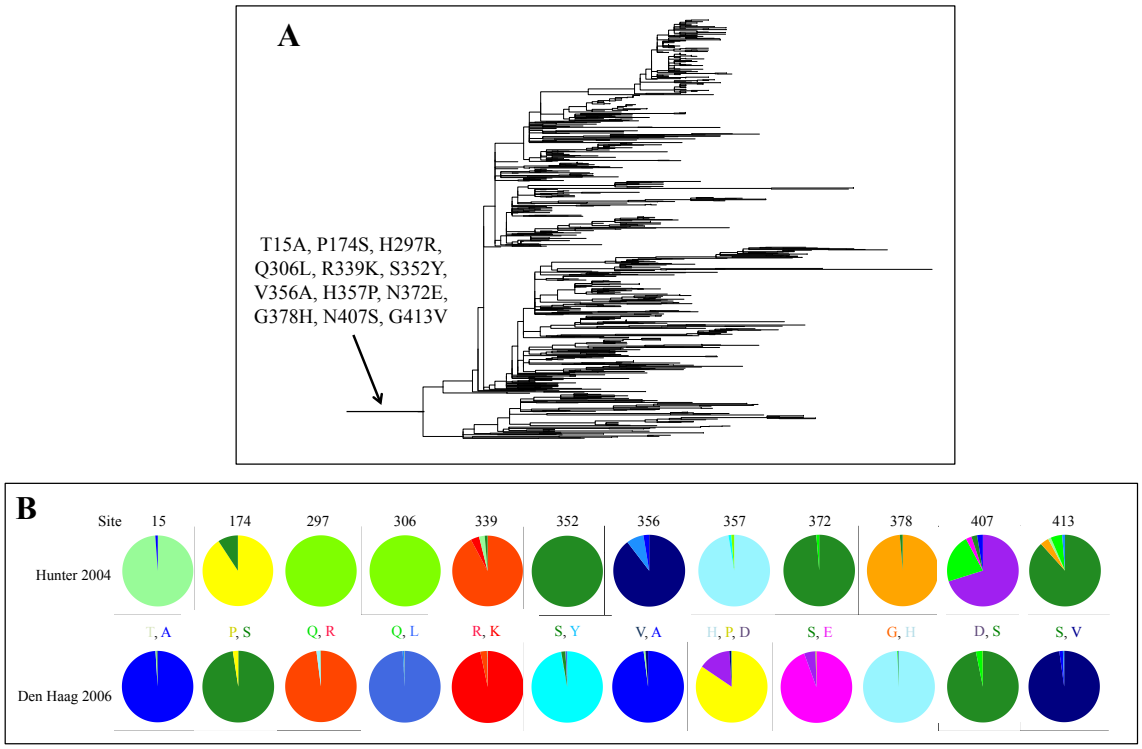


Figure 4.14: Nonsynonymous substitutions leading to the Den Haag 2006 clade. (**A**) Maximum likelihood tree of 1137 Den Haag 2006 capsid sequences. The nonsynonymous substitutions that occurred leading to the common ancestor of the Den Haag 2006 clade in more than half of the tested tree topologies are labelled. (**B**) Comparison of the amino acid distribution at each of the sites that change leading to the common ancestor of the Den Haag 2006 clade with that in the previous pandemic strain, Hunter 2004. Amino acid residues at each site are shown between the distributions and residues within similar properties are shown in similar colours.

gions. Experimental investigation of subsets of the substitutions H297R, Q306L, R339K, S352Y, V356A, H357P, N372E, G378H, N407S and G413V may determine which substitutions truly enabled the pandemic emergence of Den Haag 2006.

Hunter 2004

While it is well supported that the Hunter 2004 pandemic strain clusters with the Yerseke 2006 epidemic strain (Figure 2.1), the exact phylogenetic relationship within this cluster is not well supported (Figure 4.17). In particular, the clustering of three sequences sampled in Paraguay and South Africa and classified as Hunter 2004 viruses by the norovirus genotyping tool is uncertain. We therefore examined the Paraguay and South Africa sequences and the early changes within the Hunter 2004/Yerseke 2006 clade to determine whether antigenic changes likely occurred within the early part of this clade

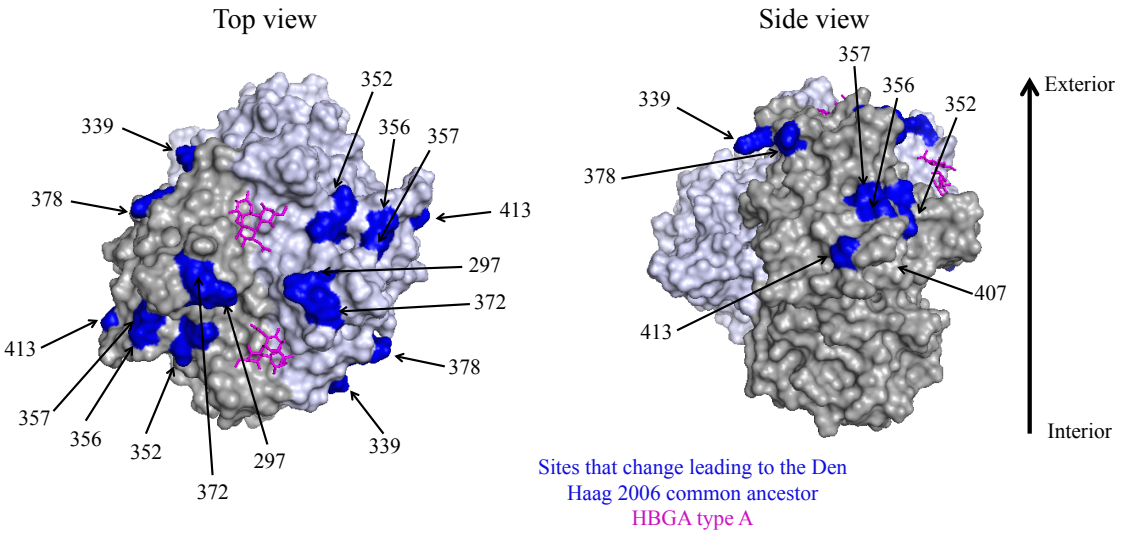


Figure 4.15: Surface location of sites that change leading to the Den Haag 2006 common ancestor. Two views are shown of the Den Haag 2006 capsid structure in which the surface-exposed sites that change leading to the common ancestor of the Den Haag 2006 strain are labelled in blue. Sites 297, 339, 352, 356, 357, 372, 378, 407 and 413 are labelled.

and therefore which node likely represented the true common ancestor of Hunter 2004.

While the Paraguay sequences were typed as Hunter 2004 viruses by the norovirus genotyping tool, they were reported as a new strain by Galeano et al because they did not clearly cluster with either the Hunter 2004 or Yerseke 2006 reference strains in their analysis (Galeano et al., 2013). Additionally, they report differences between these sequences and Hunter 2004 in blockade epitopes A and E and therefore suggest that they may be antigenically distinct from Hunter 2004 (Galeano et al., 2013). While our ancestral reconstruction results do corroborate these viruses differing in epitope E, we find no differences in epitope A compared to the common ancestor of the main Hunter 2004 clade (node 3 in Figure 4.17). Galeano et al likely obtained differences at epitope A due to comparing the Paraguay sequences against only two Hunter 2004 reference sequences. The South Africa sequence was reported as a Hunter 2004 virus by Mans et al (Mans et al., 2010). In their phylogenetic analysis, Mans et al included a single reference virus from each strain and, while the South Africa sequence clusters with the Hunter 2004 and Yerseke 2006 reference sequences in a capsid phylogeny, there is low support on the exact relationship within this clade. We observe the same in our analysis and ancestral reconstruction suggests a large number of nonsynonymous substitutions occurred leading to the South Africa sequence, including substitutions in epitope regions. Therefore

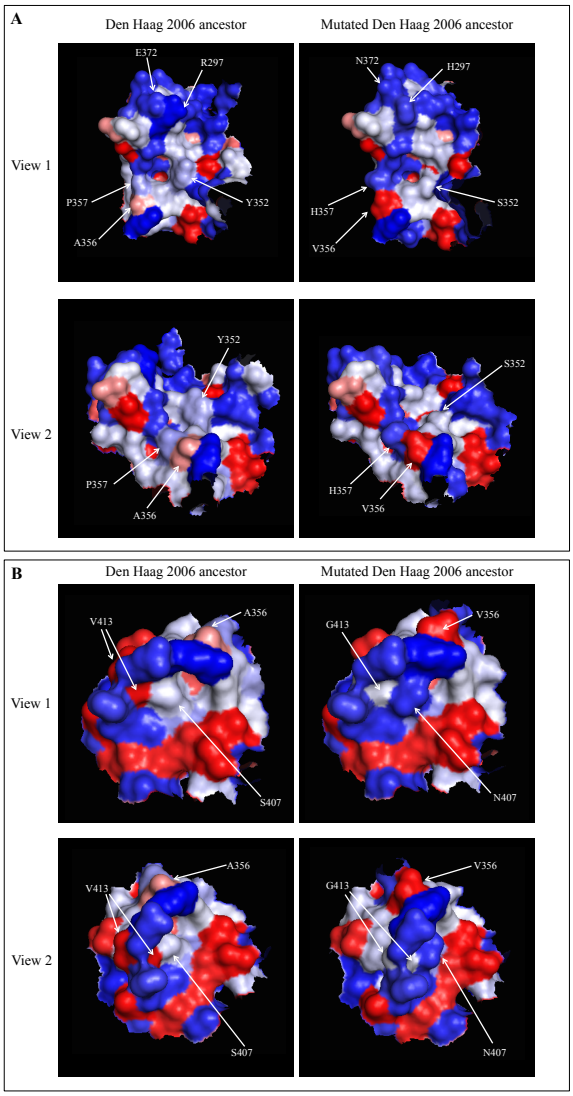


Figure 4.16: Changes in capsid structure leading to the Den Haag 2006 common ancestor. We reconstructed a homology model of the Den Haag 2006 common ancestor and the preceding ancestral sequence (here called mutated Den Haag 2006 ancestor) and examined the structural changes induced by the substitutions leading to Den Haag 2006. (**A**) Sites within 12 *Å* of either site 297 or site 352 are shown from two viewpoints. View 1 is a top view of the capsid structure, while view 2 is a side view of the capsid structure. The substitutions at sites 297 and 372 are predicted to result in a change to the structure of a ridge in blockade epitope A. The substitutions at sites 352, 356 and 357 are predicted to alter the surface structure between blockade epitopes A and E. A356 and P357 in the Den Haag 2006 common ancestor result in a smaller structure along a surface-exposed ridge compared to the V356 and H357 in the upstream ancestor. Y352 in the Den Haag 2006 common ancestor also increases the surface area of a nearby ridge compared with S352 in the upstream ancestor. (**B**) Sites within 12 *Å* of residue 407 are shown, with residues in this area forming blockade epitope E (Lindesmith et al., 2012a). Two views from the side of the capsid structure are shown. Small structural changes within blockade epitope E are predicted due to the N407S and G413V substitutions leading to the Den Haag 2006 common ancestor.

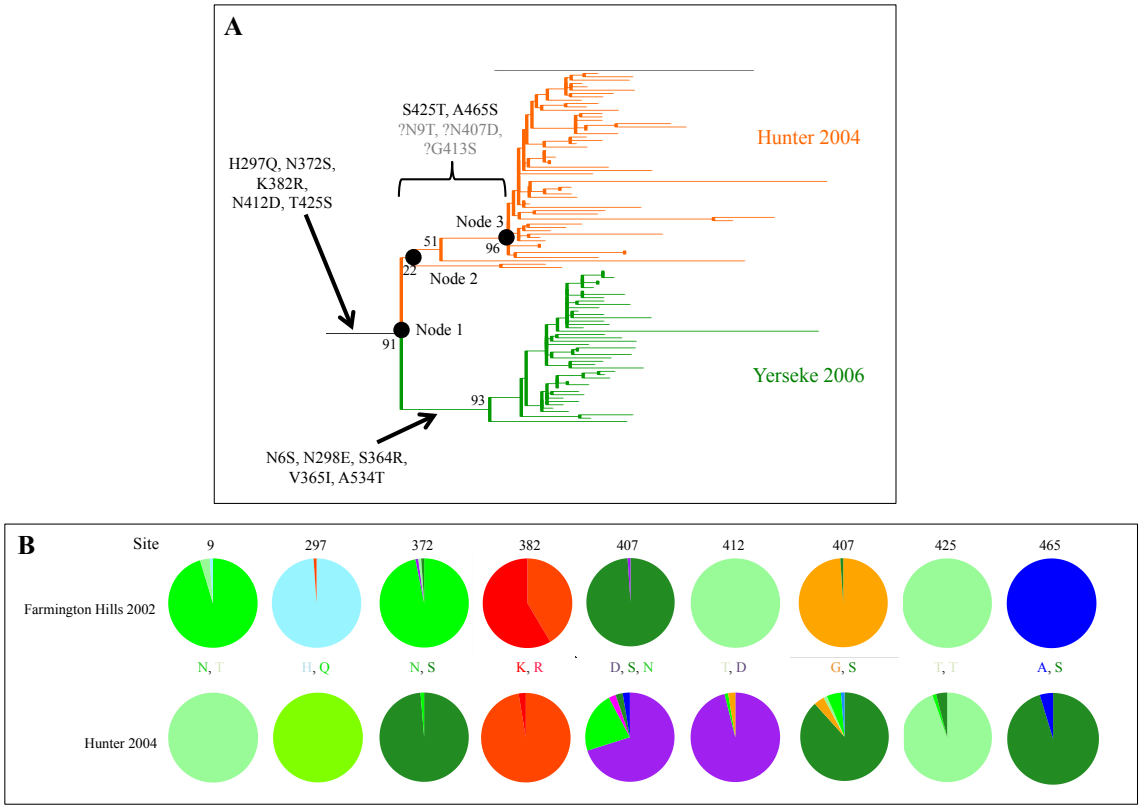


Figure 4.17: Substitutions leading to the Hunter 2004 clade. (**A**) The Hunter 2004/Yerseke 2006 clade was extracted from the maximum likelihood tree in Figure 2.1 and branches coloured by strain, as assigned by the norovirus genotyping tool: orange - Hunter 2004, green - Yerseke 2006. Bootstrap supports are shown on key nodes. We defined three nodes that may have been the common ancestor of Hunter 2004: node 1 which is the common ancestor of the Hunter 2004 and Yerseke 2006 clades, node 2 which is the common ancestor of the sequences genotyped by the norovirus genotyping tool as Hunter 2004 and node 3 which is the common ancestor of the Hunter 2004 sequences excluding the Paraguay and South Africa sequences. Nodes 1 and 3 are well supported, while node 2 is not present in most bootstrap tree topologies. The nonsynonymous substitutions leading to node 1, leading to the Yerseke 2006 common ancestor and between nodes 1 and 3 are labelled. It is well supported that the substitutions S425T and A465S occurred leading to node 3. There is lower support that the substitutions N9T, N407D and G413S occurred between node 1 and node 3. (**B**) Comparison of the amino acid residue distribution at each of the sites that may have changed leading to the common ancestor of the Hunter 2004 clade. The distribution is shown for Hunter 2004 and for the preceding pandemic strain, Farmington Hills 2002. The residues present at each site are shown between the distributions and amino acid residues with similar properties are shown in similar colours.

both the Paraguay and South Africa sequences contain substitutions in known epitope regions relative to the main Hunter 2004 clade. There is, however, no reason to assume that all substitutions within epitope regions will result in viruses with substantially different antigenic properties and future experimental studies examining the antigenicity of the Paraguay and South Africa viruses are needed to determine whether they belong to the Hunter 2004 strain. Therefore the Hunter 2004 common ancestor may have been the ancestor of the Hunter 2004 and Yerseke 2006 strains (node 1 in Figure 4.17), the common ancestor of the Hunter 2004 sequences including the Paraguay and South Africa sequences (node 2 in Figure 4.17) or the common ancestor of the Hunter 2004 sequences excluding the Paraguay and South Africa sequences (node 3 in Figure 4.17).

If node 1 was the true common ancestor of Hunter 2004, the Paraguay and South Africa sequences would be Hunter 2004 viruses and the Yerseke 2006 epidemic strain would have evolved from a Hunter 2004-like virus. It is well supported that five nonsynonymous substitutions occurred leading to node 1: H297Q, N372S, K382R, T412D and T425S. Sites 297, 372 and 412 exhibit a different distribution of amino acids in Hunter 2004 compared with that in the preceding pandemic strain, Farmington Hills 2002 (Figure 4.17). Sites 297 and 372 are located within blockade epitope A while site 412 is located within blockade epitope E (Lindesmith et al., 2012a). Homology modelling suggests the substitutions at sites 297 and 372 may have altered the structure within blockade epitope A, while the substitution at site 412 resulted in only a small difference in structure (Figure 4.18).

Node 2 is poorly supported and does not exist in seven of the eleven tested tree topologies. Where node 2 does exist, the N9T and possibly T534A substitutions occurred leading to this node. The location of these sites within the shell domain and P1 domain, respectively, suggests that substitutions at these sites are unlikely to alter viral antigenicity, suggesting that should node 2 have existed, it was likely antigenically indistinguishable from node 1.

It is well supported that the substitutions S425T and A465S occurred leading to node 3 (Figure 4.17). However, each of these sites is located towards the base of the P1 domain and is therefore unlikely to alter the viral antigenicity. Therefore it is again unlikely that there was a change in antigenicity leading to node 3 and the viruses downstream of node 3 are therefore unlikely to contain unique and conserved antigenic changes. However,

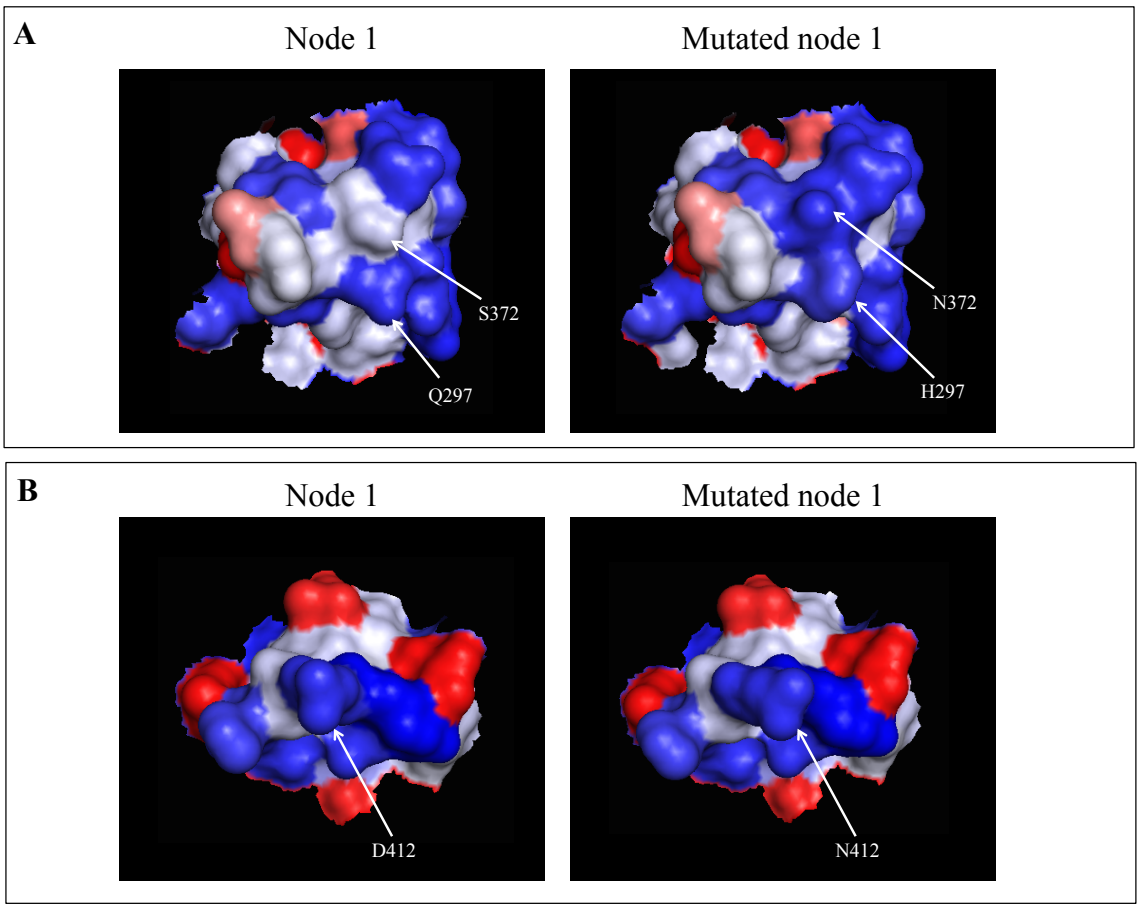


Figure 4.18: Substitutions leading to Hunter 2004 node 1. We reconstructed homology models of the sequence at node 1 in the Hunter 2004 and Yerseke 2006 clade and the upstream ancestral sequence (here called mutated node 1). Residues are coloured by their hydrophobicity based of the Kyte and Doolittle scale (Kyte and Doolittle, 1982), with blue residues being more polar and red residues being hydrophobic. (**A**) Residues within 12*Å* of site 297 are shown from a top view of the capsid. Sites 297 and 372 are labelled in each structural model and are predicted to result in a small difference in structure within epitope A. (**B**) Residues within 12*Å* of site 412 are shown from a side view of the capsid. The substitution N412D does not result in a change in residue size or hydrophobicity.

in 6 of the 11 tested tree topologies, the substitution N407D occurs between node 1 and node 3, with G413S also occurring in 4 of these 6 tree topologies. Sites 407 and 413 are both located within blockade epitope E and the substitutions at these sites occur after the branching of one or both of the Paraguay and South Africa sequences, but not along the branch leading to node 3 (Figure 4.17). It is therefore possible that there was a change in antigenicity between node 1 and node 3.

We demonstrated in chapter 3 that node 1 most likely occurred in 2001 (Table 3.4). The Paraguay and South Africa sequences were not included in the dataset we used in chapter 3 and the Hunter 2004 common ancestor in Figure 3.1 is therefore node 3, which we inferred to occur in late 2002 (Table 3.3). Therefore at the onset of the Hunter 2004 pandemic the lineages leading Yerseke 2006, the Paraguay sequences and the South Africa sequence were already present. However, viruses in the Yerseke 2006 lineage were not detected as contributing to the Hunter 2004 pandemic and the Paraguay and South Africa lineages have each been sampled in a single study, with the South Africa sequence being collected in late 2008, two years after the end of the Hunter 2004 pandemic. While it is possible that different clades within a pandemic strain may be sampled at different frequencies during the pandemic, it is very unlikely that the vast majority of sampled viruses would come from a single clade in the pandemic strain with no viruses being sampled from another clade within the pandemic. Therefore it is very unlikely that the Yerseke 2006 lineage, Paraguay lineage or South Africa lineage contributed greatly to the Hunter 2004 pandemic and we therefore suggest that these lineages do not have one or more of the defining features important for the pandemic emergence of Hunter 2004. We therefore hypothesise that the viruses at node 1 and node 3 differed antigenically. Future experimental work examining the antigenicity of the reconstructed sequence at each node, the influence of the substitutions leading to node 1 and the impact of N407D and G413S substitutions that may have occurred between nodes 1 and 3 is required to verify which node is the true common ancestor of Hunter 2004 and which substitutions likely enabled the pandemic emergence of this strain.

Farmington Hills 2002

There are 9 nonsynonymous substitutions that occurred leading to the Farmington Hills 2002 common ancestor in most tested tree topologies: S9N, D298N, K329R, S355D,

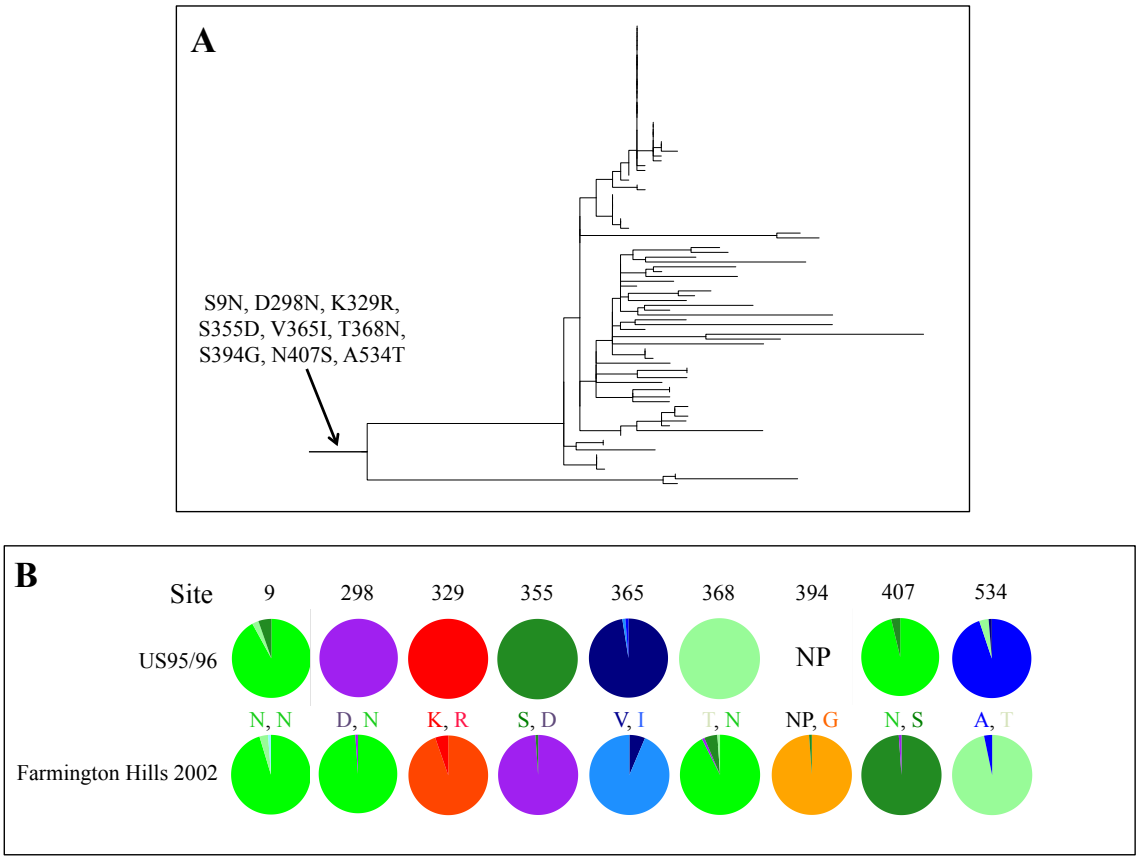
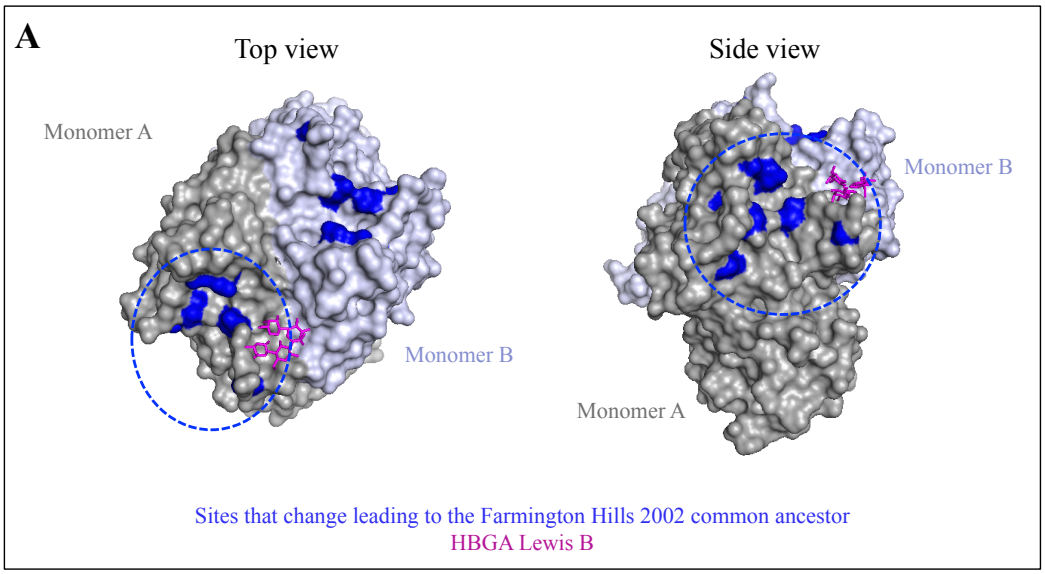


Figure 4.19: Nonsynonymous substitutions leading to the Farmington Hills 2002 clade. (**A**) Maximum likelihood tree of 96 Farmington Hills 2002 capsid sequences. Each of the nonsynonymous substitutions that occurred leading to the Farmington Hills 2002 clade is labelled. (**B**) Comparison of the amino acid residue distribution between Farmington Hills 2002 and the preceding pandemic strain, US95/96 at each site that change leading to the common ancestor of the Farmington Hills 2002 clade. The residues at each site are shown between the distributions and residues with similar properties are shown in similar colours. NP - not present, site 394 was not present in the US95/96 strain as it arose in an insertion event after divergence from US95/96. However, this insertion is present in all other pandemic and major epidemic strains post-US95/96, with the exception of Kaiso 2003.

V365I, T368N, S394G, N407S and A534T (Figure 4.19). Site 9 is the only one of these sites to not exhibit a different amino acid residue in Farmington Hills 2002 compared with the preceding pandemic strain, US95/96 (Figure 4.19).

Sites 9 and 534 are located within the shell domain and close to the base of the P1 domain, respectively, and are therefore unlikely to alter viral antigenicity. Each of the remaining sites that change leading to the Farmington Hills 2002 common ancestor are found close together within the P2 domain (Figure 4.20). Homology models of the Farmington Hills 2002 common ancestor and the upstream node suggest that these substitutions may have altered the structure of blockade epitopes A, D and E (Figure 4.20).



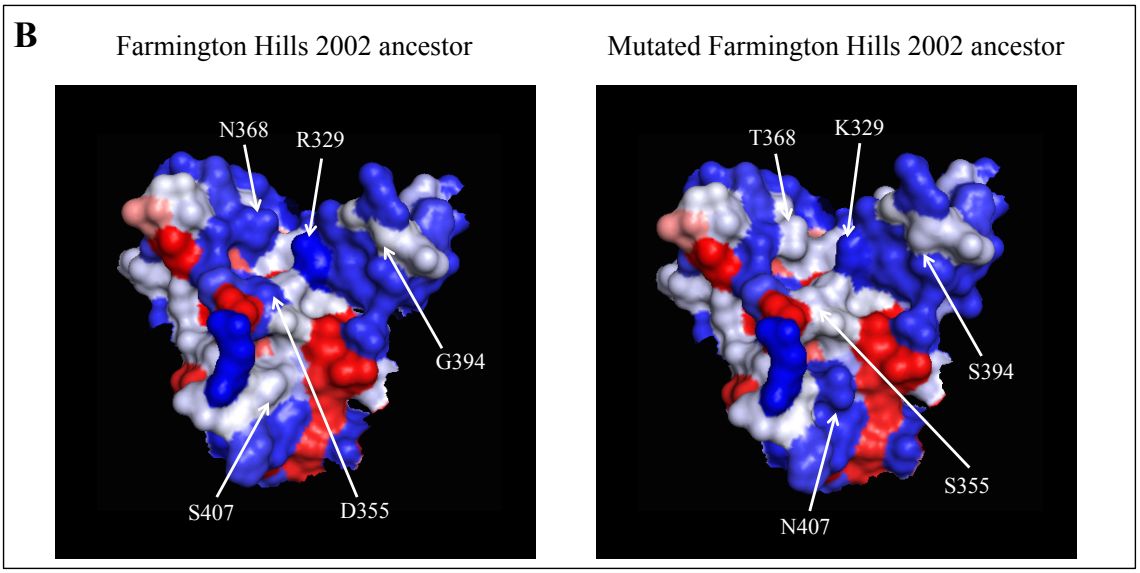


Figure 4.20: Changes in capsid structure leading to the Farmington Hills 2002 clade. (**A**) The nonsynonymous substitutions leading to the Farmington Hills 2002 clade occurred close together within the P2 domain. Sites 298, 329, 355, 365, 368, 394 and 407 are labelled in blue. The sites are shown on dimeric homology model of the Farmington Hills 2002 common ancestor, with the monomers shown in different shades and the structure is shown from a top view and a side view. The blue circle shows the location of the sites that change leading to the Farmington Hills 2002 clade. The HBGA Lewis B is shown in magenta. (**B**) We reconstructed P domain homology models of the Farmington Hills 2002 common ancestor and the upstream ancestor (here referred to as mutated Farmington Hills 2002 ancestor), which differ at sites 298, 329, 355, 365, 368, 394 and 407. Residues within 12Å of site 298, site 394 or site 407 are shown and are coloured by hydrophobicity based on the Kyte and Doolittle scale (Kyte and Doolittle, 1982). The substitutions at sites 355, 368, 394 and 407 are predicted to result in changes in surface structure.

It has previously been suggested that an insertion at site 394 in epitope D resulted in a large change in antigenicity between US95/96 and Farmington Hills 2002. However, this insertion is not unique to Farmington Hills 2002 and the common ancestor of viruses with this insertion occurred in late 1993 (95% HPD early 1991-early 1996) (Figure 2.2). Therefore at the onset of the Farmington Hills pandemic in 2002 a large number of additional lineages with this insertion were also present: the Osaka 2007 lineage, the Asia 2003 lineage, the Den Haag 2006 lineage, the Hunter 2004 lineage, the Yerseke 2006 lineage and the Apeldoorn lineage (Figure 3.1). Therefore this insertion is unlikely to have been the key change enabling pandemic emergence of Farmington Hills 2002. Lindesmith et al demonstrated that Farmington Hills 2002 has a different antigenic profile in blockade epitope E compared with US95/96 and Den Haag 2006 viruses (Lindesmith et al., 2012b). They mapped this difference to residues 407, 412 and 413 and demonstrated that by engineering these residues into the other strains, they could recapitulate the Farmington Hills 2002 antigenic phenotype. Our analysis suggests that site 407 is the only one of these three sites that changed leading to the Farmington Hills 2002 common ancestor.

Our results identify sites within several surface-exposed regions of the P2 domain that may have been important for the pandemic emergence of Farmington Hills 2002 and highlight the substitutions D298N, K329R, S355D, V365I, T368N, S394G and N407S for experimental validation of their effect on viral antigenicity.

4.4.5 Comparison of substitutions leading to each pandemic GII.4 strain

The nonsynonymous substitutions leading to each pandemic GII.4 strain are predominantly strain-specific (Figure 4.21) with no evidence for a particular site that consistently enables the emergence of new pandemic strains. The majority of these substitutions are located within the P2 domain, with only a small number of sites located in the S and P1 domains (Figure 4.21). There are four sites that changed leading to more than one pandemic clade: 297, 368, 372 and 407. Sites 297, 368 and 372 are located within blockade epitope A and site 407 is located within blockade epitope E (Lindesmith et al., 2012a).

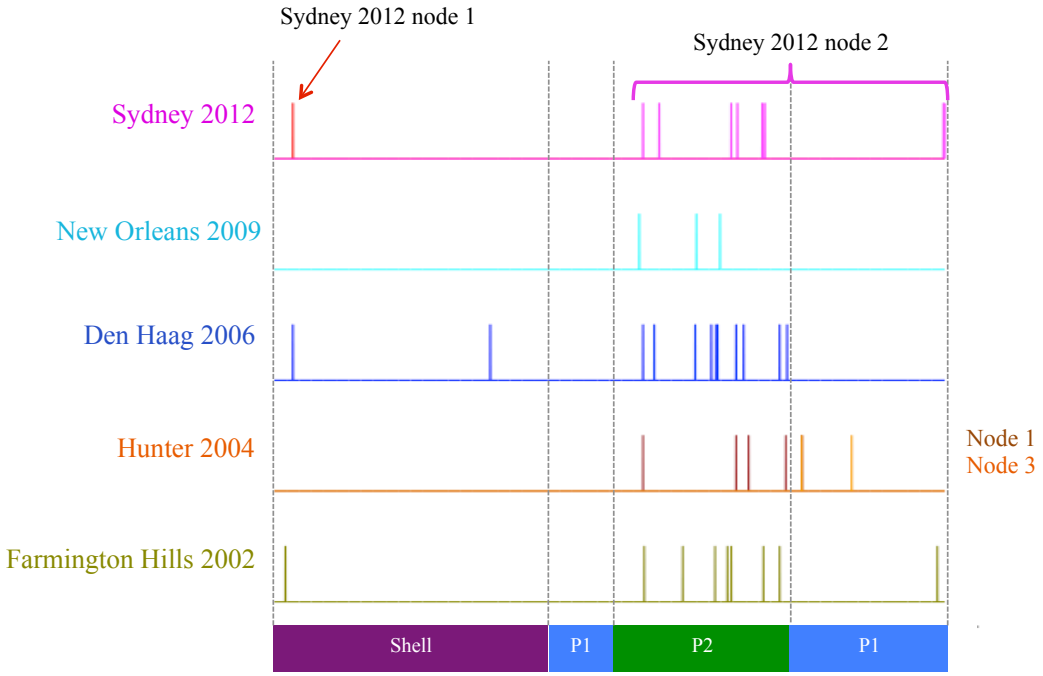


Figure 4.21: Location of nonsynonymous capsid substitutions leading to pandemic clades. The location within the capsid is shown for each of the nonsynonymous substitutions that occurred leading to the Farmington Hills 2002, Hunter 2004, Den Haag 2006, New Orleans 2009 and Sydney 2012 clades. For Hunter 2004, the substitutions leading to node 1 are shown in brown and the substitutions leading to node 3 are shown in orange. For Sydney 2012, the substitution leading to node 1 is shown in red and the substitutions leading to node 2 are shown in magenta. The vertical grey dashed lines demonstrate the domain boundaries.

4.4.6 Validation of substitutions important for the pandemic emergence of Sydney 2012

Under the three stage process of strain emergence introduced in chapter 3 (Figure 3.6), we expect that the genetic changes required for pandemic spread are acquired by the common ancestor of the strain. To verify whether this is the case for each of the pandemic strains and to determine which nonsynonymous substitutions were likely the most important for pandemic emergence, we will employ reference sets of monoclonal antibodies and polyclonal sera raised against individual GII.4 pandemic strains. We have initially designed three virus like particles (VLPs) to test the antigenic properties of ancestral Sydney 2012 viruses and to test our hypothesis that one or more of the substitutions at sites 310, 368 and 373 were key for enabling the pandemic emergence of Sydney 2012. The first two VLPs contain the inferred ancestral sequences at Sydney 2012 node 1 and node 2, which differ at eight sites (Figure 4.10). The third VLP (VLP3) contains the node 1 se-

quence with the three substitutions we hypothesise to have enabled pandemic emergence of Sydney 2012: S310N, A368E and N373H (Figure 4.11). We will test the reactivity of each of these VLPs with a panel of monoclonal antibodies and convalescent polyclonal sera raised against New Orleans 2009 and Sydney 2012 and compare the binding profiles of these three VLPs with a reference set of New Orleans 2009 and Sydney 2012 VLPs. Under our hypothesis, we would expect the node 2 VLP and VLP3 to exhibit similar antigenic properties to the reference Sydney 2012 VLPs and for the node 2 VLP and VLP3 to be antigenically distinct from the reference New Orleans 2009 VLPs. We also hypothesise that the node 2 VLP and VLP3 will be antigenically distinct from the node 1 VLP. This will enable us to determine whether the node 2 virus had acquired a 'Sydney 2012-like' antigenic profile, whether additional important antigenic changes were acquired between node 1 and node 2, whether the node 1 and node 2 viruses were antigenically distinct from New Orleans 2009 viruses and whether one or more of the substitutions at sites 310, 368 and 373 resulted in a large change in antigenicity.

Importantly, the panel of monoclonal antibodies were raised against individual VLPs from New Orleans 2009 and Sydney 2012. Comparison of the node 1 and node 2 sequences with the reference VLP (VLPref) used to raise the Sydney 2012 monoclonal antibodies identified sequence differences, namely I119V, I145V, I231V, S/N310D, A340T, N/H373R and N/S393G (Table 4.2). The VLPref was created from the first Sydney 2012 virus to be identified (accession number JX459908.1), but this virus is not representative of the complete Sydney 2012 clade at a number of sites. For example, this sequence has residue D at site 310, which is very rare within the Sydney 2012 clade (Figure 4.11). Table 4.2 summarises the differences between the VLPref, node 1 VLP and node 2 VLP. The sites that have a unique residue within the VLPref compared with node 1 and node 2 may influence the ability of the monoclonal antibodies to recognise the node 1 and node 2 VLPs. Unfortunately, the reference Sydney 2012 VLP differs from the node 1 and node 2 VLPs at two of the three sites of interest, 310 and 373. It is therefore possible that the monoclonal antibodies will not give information on the importance of the substitutions at these sites. However, the panel of convalescent polyclonal sera should give an accurate reflection of the antigenicity of the respective VLPs and on the combined importance of the substitutions at sites 310, 368 and 373.

While we have initially synthesised VLPs based on the Sydney 2012 ancestral se-

Amino acid site	Reference Sydney 2012 VLP residue	Node 1 residue	Node 2 residue	VLP3 residue
119	V	I	I	I
145	V	I	I	I
231	V	I	I	I
297	R	H	R	H
310	D	S	N	N
340	T	A	A	A
368	E	A	E	E
373	R	N	H	H
393	G	N	S	N
395	T	A	T	A
539	A	A	V	A
540	V	L	V	L

Table 4.2: Variable sites between Sydney 2012 VLPs. Each of the variable sites between any of the Sydney 2012 VLPs is shown. All other sites are constant between the VLPs. The reference Sydney 2012 VLP was used in construction of the Sydney 2012 reference monoclonal antibodies. The node 1 and node 2 VLPs differ at sites 297, 310, 368, 373, 393, 395, 539 and 540. VLP3 has the same sequence as the node 1 VLP with the exception of sites 310, 368 and 373. Sites 297, 310, 368, 373, 393, 395, 539 and 540 change between node 1 and node 2, while the other sites are included here because they change between the reference Sydney 2012 VLP sequence and our reconstructed sequences. Sites 310, 368 and 373 are shown in bold as these are the sites we hypothesise to have enabled the pandemic emergence of Sydney 2012.

quences and inferred substitutions, we will carry out the same process for the other pandemic GII.4 strains using the ancestral sequences and substitutions identified above.

4.4.7 Strains accumulate diversity during the diversification phase of strain emergence

We next characterised the diveristy at the amino acid level within the capsid of each of the pandemic GII.4 strains. We could accurately characterise the diversity in the three most recent pandemic strains, Den Haag 2006, New Orleans 2009 and Sydney 2012, due to the large number of capsid sequences available for these strains. We identified 19, 13 and 15 amino acid sites that exhibit a high degree of variability within Den Haag 2006, New Orleans 2009 and Sydney 2012, respectively (Figures 4.22, S4.7-S4.9, Table 4.3). Identifying variable sites within US95/96, Farmington Hills 2002 and Hunter 2004 was more challenging due to the smaller number of capsid sequences available for these

strains. The total branch length in our nucleotide maximum likelihood trees for US95/96, Farmington Hills 2002 and Hunter 2004 was far smaller than that for Den Haag 2006, New Orleans 2009 and Sydney 2012, suggesting that our sample sets capture less diversity for these earlier pandemic strains. Correspondingly, we find a correlation between the total nucleotide branch length in the strain phylogenetic tree and the total amino acid entropy within the strain (Figure 4.22). To reduce the potential for false positives, we took sites as variable in US95/96, Farmington Hills 2002 and Hunter 2004 only if they exhibited a greater degree of variability. We identify 3, 2 and 8 variable sites in US95/96, Farmington Hills 2002 and Hunter 2004, respectively (Figures 4.22, S4.4-S4.6, Table 4.3). As would be expected, most of the variable sites are located within the P2 domain (Figure 4.22), consistent with this region being able to accomodate a greater degree of structural variability compared with the shell and P1 domains and the persistent targeting of this region by host immunity.

The intra-strain variable sites often coincide with epitope positions (Figure 4.23, Table 4.3). 11 of the 12 sites in blockade epitopes A, D and E are variable within at least one of the pandemic strains, with site 394 in epitope D being the exception. Additionally, each of the sites within putative epitopes B and C is highly variable within at least one pandemic strain. It was previously suggested that the surface epitopes may include additional residues within 8 $\AA$ of the proposed epitope positions (Lindesmith et al., 2012a). The intra-strain variable sites include additional positions within 8 $\AA$ of blockade epitopes A (site 373), D (site 396) and E (sites 356 and 414) as well as putative epitope C (sites 339, 341 and 377).

Several of the intra-strain variable sites coincide with the HBGA binding region (Figure 4.24, Table 4.3). The sites in the HBGA binding region can be divided into those that are required for binding to all HBGAs and those that are required for binding to only specific HBGAs (Singh et al., 2015). The residues required for binding to all HBGAs (T344, R345, D374, G443 and Y444) are conserved in all GII.4 strains and throughout the GII.4 clade. However, the variable sites 393 and 395 are located within a loop that interacts with specific HBGAs (Singh et al., 2015) and there are multiple sites that exhibit high intra-strain variability and are located close to the HBGA-specific binding site (Figure 4.24, Table 4.3). That the intra-strain variable sites often coincide with sites within or close to epitopes and the HBGA-binding region indicates that this variability has the

Site	Domain	Epitope site	HBGA binding site	Close to epitope site	Close to HBGA binding site	Type of variation
US9596						
47	Shell	No	No	No	No	Large clade
333	P2	Epitope B	No	No additional epitope	Site 441	Multiple changes
393	P2	Epitope D	Yes	No additional epitope	N/A	Multiple changes
Farmington Hills 2002						
382	P2	Epitope B	No	No additional epitope	No	Multiple changes
395	P2	Epitope D	Yes	No additional epitope	N/A	Multiple changes
Hunter 2004						
174	Shell	No	No	No	No	Multiple changes
296	P2	Epitope A	No	No additional epitope	296	Multiple changes
329	P2	No	No	No	No	Multiple changes
339	P2	No	No	Epitope C	Sites 343, 344, 345	Multiple changes
340	P2	Epitope C	No	No additional epitopes	Site 343	Multiple changes
356	P2	No	No	Epitope E	No	Multiple changes
407	P2	Epitope E	No	No additional epitopes	No	Multiple changes
413	P2	Epitope E	No	No additional epitopes	No	Multiple changes
Den Haag 2006						
6	Shell	No	No	No	No	Multiple changes
9	Shell	No	No	No	No	Multiple changes
23	Shell	No	No	No	No	Multiple changes
45	Shell	No	No	No	No	Large Clade
171	Shell	No	No	No	No	Large clade
193	Shell	No	No	No	No	Large clade
255	P1	No	No	No	No	Large clade
298	P2	Epitope A	No	No additional epitope	No	Multiple changes
333	P2	Epitope B	No	No additional epitope	Site 441	Multiple changes
340	P2	Epitope C	No	No additional epitopes	Site 343	Multiple changes
357	P2	No	No	No	No	Multiple changes
368	P2	Epitope A	No	No additional epitopes	No	Multiple changes
372	P2	Epitope A	No	No additional epitopes	Site 374	Multiple changes
377	P2	No	No	Epitope C	No	Multiple changes
393	P2	Epitope D	Yes	No additional epitopes	N/A	Multiple changes
412	P2	Eptiope E	No	No additional epitopes	No	Multiple changes
414	P2	No	No	Epitope E	No	Multiple changes
425	P1	No	No	No	No	Large clade
539	P1	No	No	No	No	Multiple changes
New Orleans 2009						
54	Shell	No	No	No	No	Large clade
174	Shell	No	No	No	No	Multiple changes
231	P1	No	No	No	No	Multiple changes
294	P2	Epitope A	No	No additional epitopes	No	Multiple changes
297	P2	Epitope A	No	No additional epitopes	No	Multiple changes
339	P2	No	No	Epitope C	Sites 343, 344, 345	Multiple changes
341	P2	No	No	Epitope C	Sites 343, 344, 345, 444	Multiple changes
372	P2	Epitope A	No	No additional epitopes	Site 374	Multiple changes
376	P2	Epitope C	No	No additional epitopes	Sites 345, 374	Multiple changes
377	P2	No	No	Epitope C	No	Large clade
393	P2	Epitope D	Yes	No additional epitopes	N/A	Multiple changes
396	P2	No	Yes	Epitope D	N/A	Large clade
413	P2	Epitope E	No	No additional epitope	No	Multiple changes
Sydney 2012						
119	Shell	No	No	No	No	Multiple changes
145	Shell	No	No	No	No	Multiple changes
174	Shell	No	No	No	No	Multiple changes
231	P1	No	No	No	No	Multiple changes
297	P2	Epitope A	No	No additional epitopes	No	Multiple changes
309	P2	No	No	Site 310	No	Large clade
333	P2	Epitope B	No	No additional epitopes	Site 441	Multiple changes
340	P2	Epitope C	No	No additional epitopes	Site 343	Multiple changes
372	P2	Epitope A	No	No additional epitopes	Site 374	Multiple changes
373	P2	Epitope A[a]	No	No additional epitopes	Sites 345, 374	Multiple changes
377	P2	No	No	No	No	Multiple changes
393	P2	Epitope D	Yes	No additional epitopes	N/A	Multiple changes
414	P2	No	No	Epitope E	No	Large clade
539	P1	No	No	No	No	Large clade
540	P1	No	No	No	No	Large clade

Table 4.3: Variable sites within each pandemic GII.4 strain. Each of the sites we identify as exhibiting high variability within a pandemic GII.4 strain is shown. The capsid domains were defined as in Cao et al. (2007). Epitopes were defined as those regions identified in Lindesmith et al. (2012a), with epitopes A, D and E being blockade epitopes and epitopes B and C being putative epitopes. The HBGA-binding regions were defined as those regions identified by Cao et al. (2007) and Singh et al. (2015). We classified sites as being close to an epitope or HBGA-binding region if the residue at the site is located within 8Å of one or more residues within an epitope or HBGA-binding region (Lindesmith et al., 2012a) in the Sydney 2012 P domain structure 4WZT (Singh et al., 2015). The nature of the variation at a site was identified using the coloured trees method. Sites classified as 'large clade' exhibit different amino acids within different large clades within the tree while sites classified as 'multiple changes' change on multiple occassions within the strain. Sites that change leading to multiple large clades are classified as multiple changes. Coloured trees are shown for each site in Figures S4.4-S4.9. [a] Site 373 was not included in epitope A in Lindesmith et al. (2012a) but was later included in epitope A by Debbink et al. (2013)

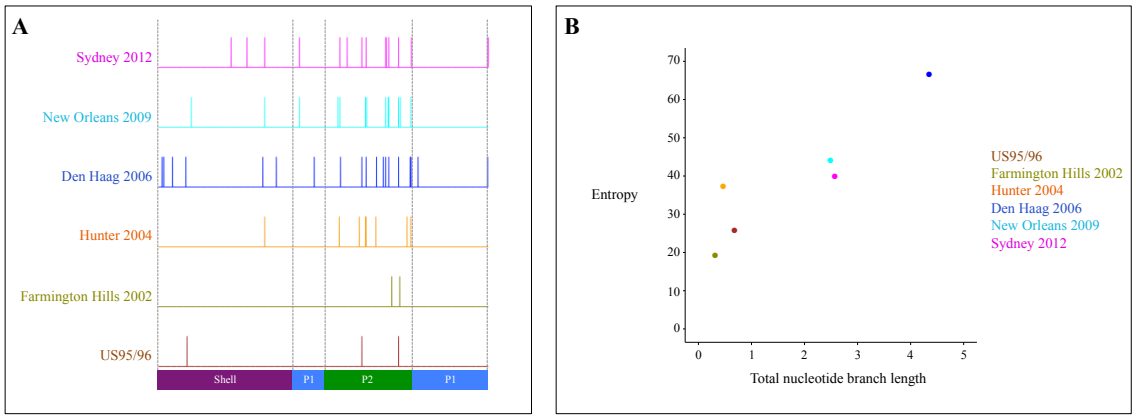


Figure 4.22: Highly diverse sites within individual GII.4 pandemic strains. (**A**) The location within the capsid sequence of diverse amino acid sites is shown for each pandemic strain. The vertical grey dashed lines are the domain boundaries. (**B**) We reconstructed a maximum likelihood phylogenetic tree for each of the pandemic GII.4 strains individually. The total nucleotide branch length within the maximum likelihood tree is plotted against the total amino acid entropy within the strain. There is a correlation, indicating that as nucleotide diversity increases, amino acid diversity increases.

Number of pandemic strains	Sites
1	6, 9, 23, 45, 47, 54, 119, 145, 171, 193, 255, 294, 296, 298, 309, 329, 341, 356, 357, 368, 373, 376, 382, 395, 396, 407, 412, 425, 540
2	231, 297, 339, 413, 414, 539
3	174, 333, 340, 372, 377
4	393

Table 4.4: Summary of variable sites within pandemic GII.4 strains. Each of the sites that is highly variable in one or more pandemic GII.4 strains is shown. The number of pandemic strains is the number of pandemic strains that site is variable in.

potential to alter the viral phenotype through altering antigenicity of the HBGA-binding profile.

The variation at most of the highly variable sites in New Orleans 2009 and Sydney 2012 began to be accumulated prior to the onset of their respective pandemics (Figures 4.25, 4.26).

The variable sites within a strain are typically strain-specific, with only a small number of sites being variable in more than one strain (Table 2.4).

4.5 Discussion

In chapter 3 we presented a three stage model of pandemic strain emergence where the strain first acquires the genetic changes that will enable it to emerge pandemically in the

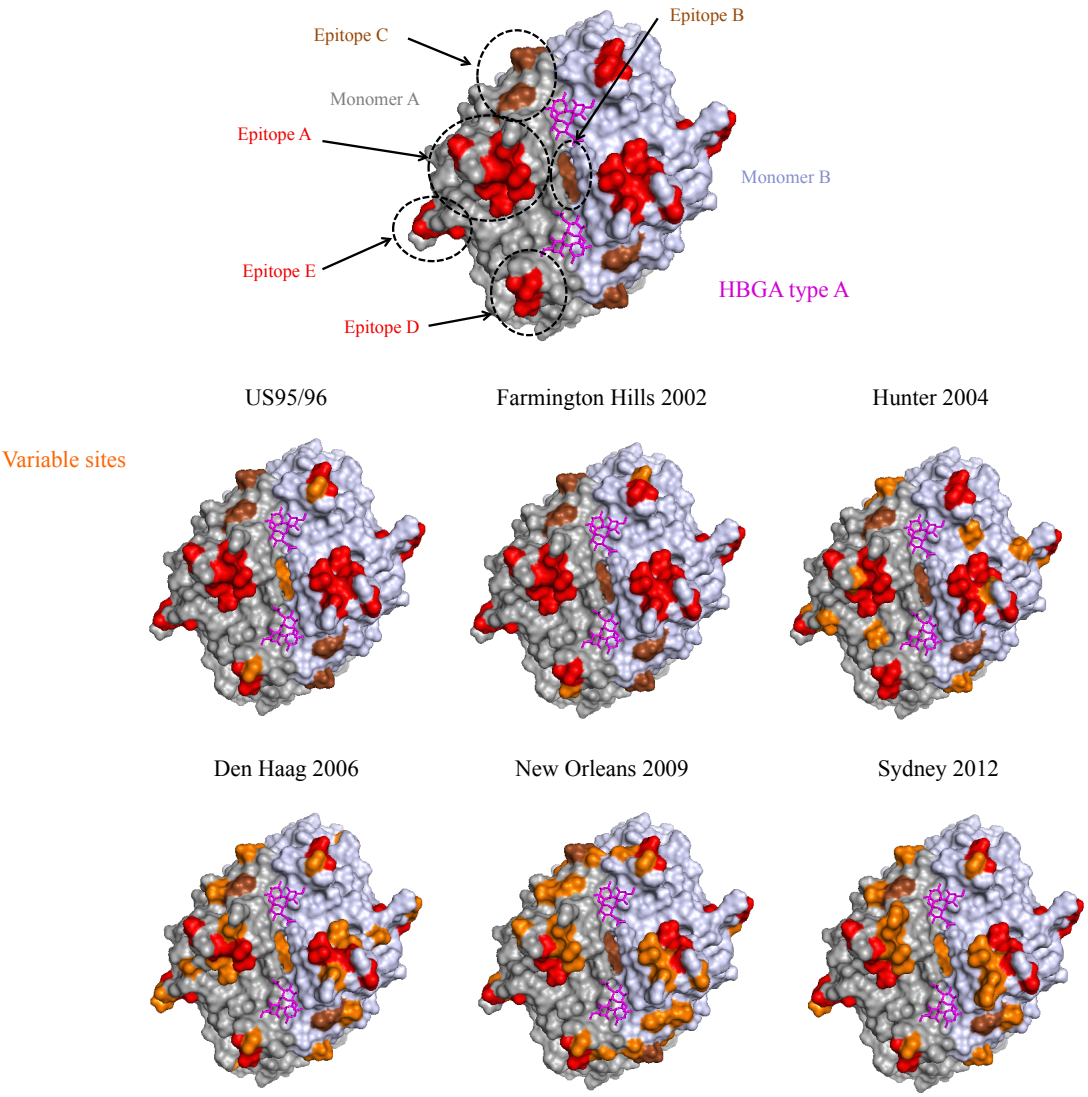


Figure 4.23: Variable sites are often located within epitope regions. A top view of the GII.4 capsid is shown based on the Sydney 2012 P domain dimer structure 4WZT (Singh et al., 2015). The monomers are labelled in different shades, with one monomer labelled in grey and the other labelled in blue-white. Each site in blockade epitopes A, D and E is shown in red and each site in putative epitopes B and C is shown in brown. The sites within each epitope region are the coloured residues enclosed within the respective dashed oval. The dashed ovals are only shown in monomer A for clarity, but the epitopes exist in each monomer. The HBGA type A bound to the Sydney 2012 P domain in structure 4WZT is shown in magenta. The same capsid view is shown with the variable sites within each pandemic strain labelled in orange.

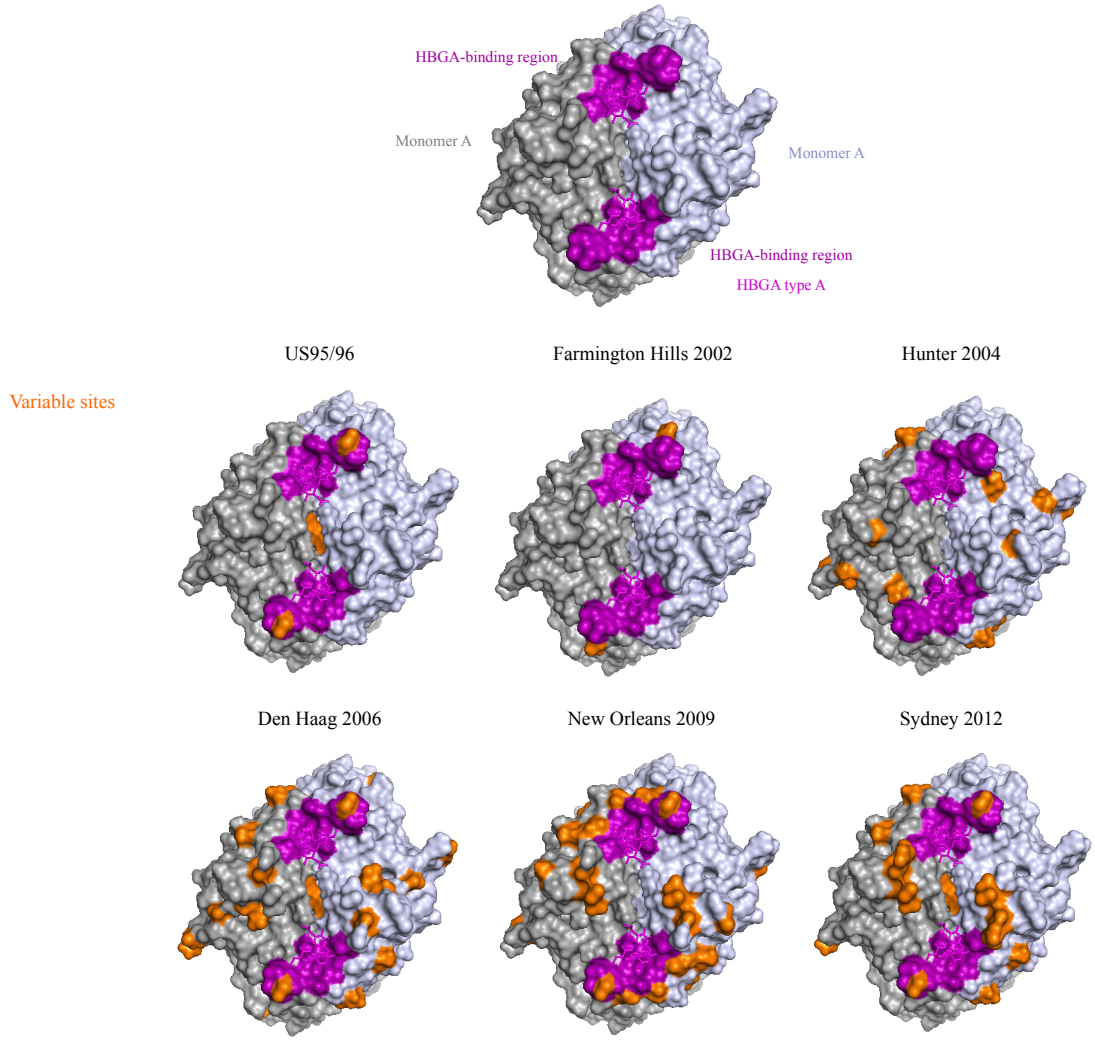


Figure 4.24: Variable sites close to and within the HBGA-binding region. A top view of the GII.4 capsid is shown based on the Sydney 2012 P domain dimer structure 4WZT (Singh et al., 2015). The monomers are labelled in different shades, with one monomer labelled in grey and the other labelled in blue-white. Each site in the HBGA-binding region is labelled in purple. The HBGA type A bound to the Sydney 2012 P domain in structure 4WZT is shown in magenta. The same capsid view is shown with the variable sites within each pandemic strain labelled in orange.

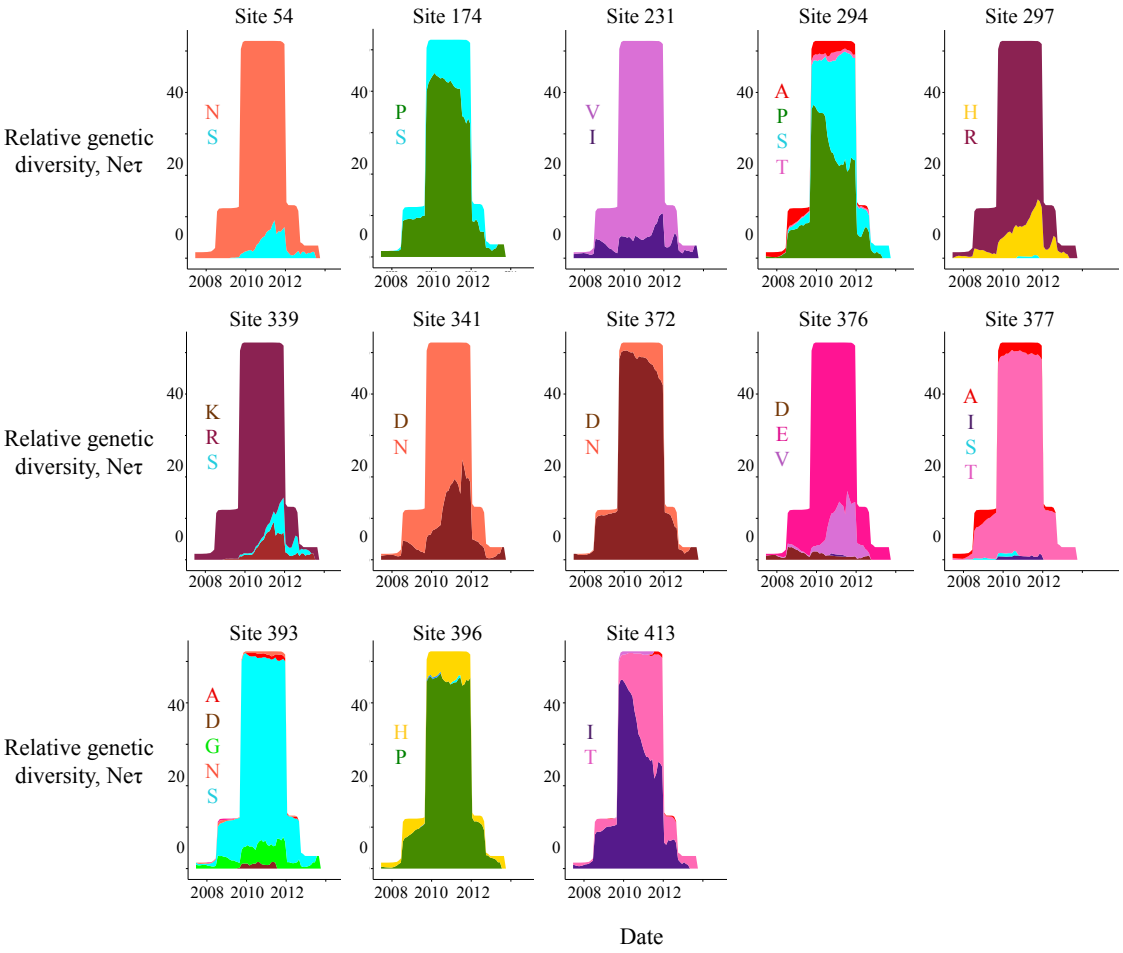


Figure 4.25: Variability in New Orleans 2009 began to accumulate prior to pandemic emergence. We carried out ancestral reconstruction on a temporally resolved New Orleans 2009 phylogenetic tree to determine how amino acid change was accumulated through time. Plotted here for each site that exhibits high variability in New Orleans 2009 is the proportion of each residue at that site through time, scaled to the inferred population history in the form of a Bayesian skyline plot. The increase in Neτ indicates the onset of the New Orleans 2009 pandemic. The residues at each site are shown next to the respective plot. The residue proportions are shown as a stacked area plot. As an example, the top right panel shows the temporal history of site 297. The presence of both histidine (H) and arginine (R) at site 297 prior to the increase in Neτ indicates diversity was accumulated at this site prior to pandemic onset.

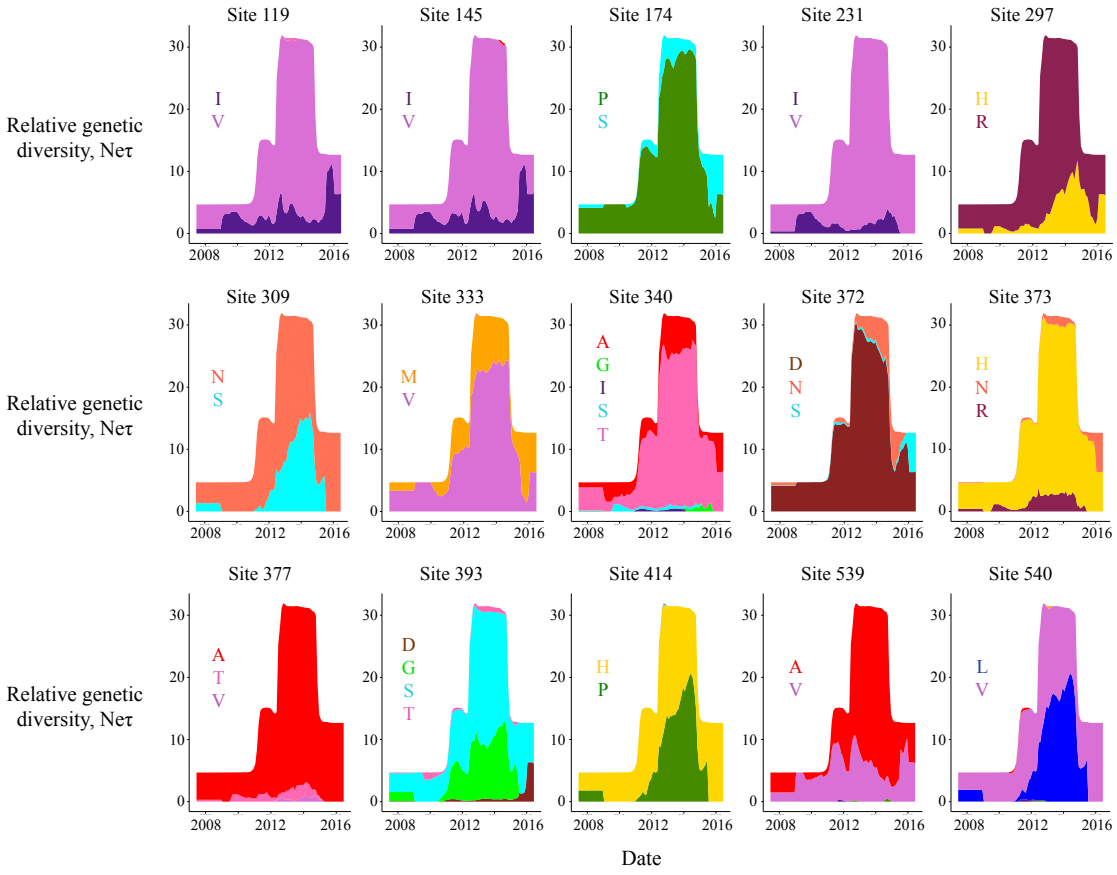


Figure 4.26: Variability in Sydney 2012 began to accumulate prior to pandemic emergence. We carried out ancestral reconstruction on a temporally resolved Sydney 2012 phylogenetic tree to determine how amino acid change was accumulated through time. Plotted here for each site that exhibits high variability in Sydney 2012 is the proportion of each residue at that site through time, scaled to the inferred population history in the form of a Bayesian skyline plot. The increase in Neτ in 2012 indicates the onset of the Sydney 2012 pandemic. The residues at each site are shown next to the respective plot. The residue proportions are shown as a stacked area plot. As an example, the top right panel shows the temporal history of site 297. The presence of both histidine (H) and arginine (R) at site 297 prior to the increase in Neτ indicates diversity was accumulated at this site prior to pandemic onset.

future and then undergoes diversification into a large number of pre-adapted lineages that emerge to cause the pandemic (Figure 3.6). We proposed that during the diversification phase the strain could either circulate within a small geographical region with spread from a single region at the onset of the pandemic or circulate widely and spread outwards from multiple geographical regions at the onset of the pandemic. The phylogeography results strongly support this second scenario, with support for New Orleans 2009 and Sydney 2012 circulating widely and consistently over several years prior to their emergence as the pandemic strain (Figures 4.3, 4.27). Therefore there is low level worldwide circulation prior to the emergence of the pandemic, explaining how New Orleans 2009 and Sydney 2012 were identified on multiple continents prior to their pandemic emergence (Figure 3.5). This strongly suggests that environmental factors do not play an important role in strain emergence, as the strain is already present on each continent prior to the pandemic and the pandemic is not caused by one sublineage from the strain that emerges from a particular geographical region. There is no support for a single continent having a special role in the early circulation of either strain (Figure 4.3). However, after the emergence of the pandemic, Asia acted as a source of viral lineages in both New Orleans 2009 and Sydney 2012, with this region exporting many more lineages than it imported (Figure 4.9). Therefore prioritising the Asia region in vaccine strategies may have a greater impact on the global epidemic compared with the impact of vaccinating other continents.

In most countries, norovirus outbreaks typically exhibit a large Winter peak followed by far fewer cases in Summer (Ahmed et al., 2013). However, the average persistence of viral lineages within each continent is much longer than one year (Figure 4.5), suggesting that the large Winter peaks are predominantly caused by viruses that have persisted within the continent during the Summer, although likely with an additional contribution from newly imported lineages. This is supported by a low overall inter-continental transmission rate (0.29 and 0.26 migration events/lineage/year for New Orleans 2009 and Sydney 2012, respectively). We do observe greatly different rates of viral migration between different pairs of continents, with a high migration rate between Europe and North America and between Asia and Oceania (Figures 4.7, 4.8). We here employed a symmetric rate matrix to model the process of inter-continental migration; studies are underway to confirm our findings using an asymmetric rate matrix that does not assume the same migration rate in each direction between each pair of continents. Migration events between

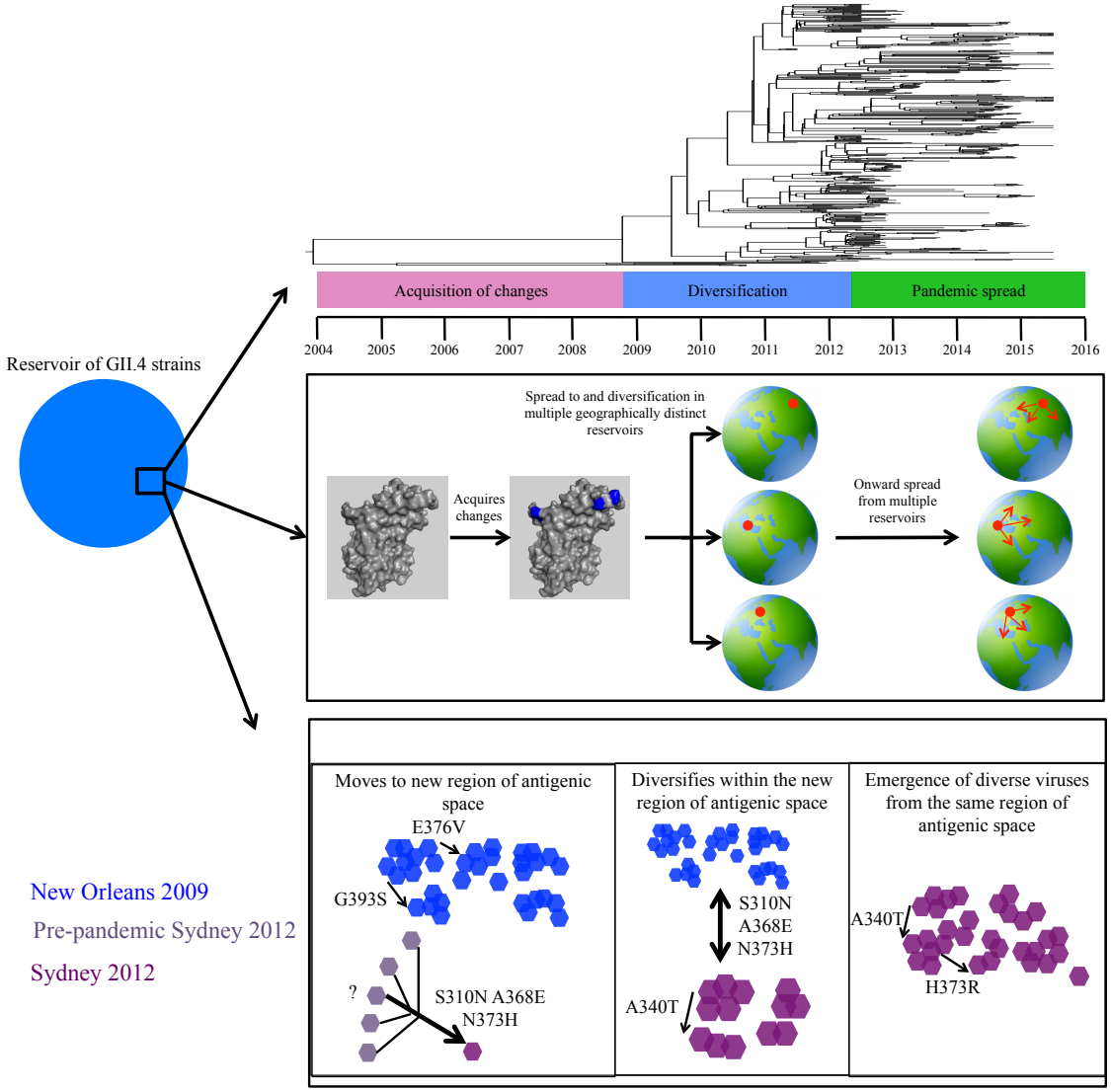


Figure 4.27: Three stage process of strain emergence including geographical spread and antigenic changes. The three stage process of strain emergence is depicted as in figure 3.6, with the strain acquiring the genetic changes that will important for it to emergence pandemically, undergoing diversification into a large number of lineages and then emerging to cause the new pandemic. The GII.4 strain that will emerge in the future is only a subset of the total reservoir of GII.4 strains. The phylogeography results strongly suggest that the strain circulates widely and consistently during the diversification phase, resulting in a period of worldwide circulation prior to pandemic emergence. The emergence of a number of closely related lineages exhibiting genetic variability suggests that these lineages share one or more phenotypic characteristics. We suggest that this characteristic is acquired during the acquisition of changes phase and moves the virus to a new region of antigenic space. It is not clear where in antigenic space the virus is prior to acquiring these changes, but after acquiring the changes, the virus is in a distinct region of antigenic space compared with the current pandemic strain. The diversification results in local movement within the antigenic space, but the lineages remain within the same general, distinct region of antigenic space. At the onset of the new pandemic, all of the lineages within the same region of the antigenic space emerge and cause the new pandemic. Here, this process is shown with reference to New Orleans 2009 as the current pandemic strain and Sydney 2012 as the next pandemic strain.

distant continents suggests that airplane travel is likely a major mode of inter-continental transmission. However, given the high frequency of flights between the majority of continents and the very high prevalence of GII.4 norovirus strains, this raises the question of why inter-continental transmission is not more frequent. While norovirus gastroenteritis is typically short lived, it is unlikely that an individual will be able to travel during the symptomatic period, which typically lasts 24-48 hours (Lopman et al., 2004a; Patel et al., 2009). Individuals do, however, shed virus in faeces for up to several months after resolution of symptoms and there is evidence of frequent asymptomatic infection (Teunis et al., 2015). There is currently no clear evidence as to whether asymptomatically infected patients may be infectious (Teunis et al., 2015), although transmission is likely to occur more frequently during symptomatic infection due to the expulsion of viral particles into the environment. Therefore we suggest that the low inter-continental transmission rate is the result of patients being unlikely to travel during the period at which they are most infectious. Inter-continental migration events likely occur through patients who have either resolved norovirus symptoms or exhibited an asymptomatic infection.

Inter-continental transmission is unlikely to be necessary for a pandemic strain to persist due to each continent containing a large susceptible population. Current prevalence estimates suggest that roughly one in ten individuals worldwide are infected with norovirus annually (Kirk et al., 2015; Pires et al., 2015; Lopman et al., 2016), with between 60 and 80% of these cases being caused by the dominant GII.4 strain (Kroneman et al., 2008). While individuals within the population are likely to be differentially susceptible to different pandemic strains through their HBGA expression (Lindesmith et al., 2008), the dominance of each pandemic strain for a typical period of 2-3 years suggests that even towards the end of the pandemic period, a large number of susceptible individuals will remain within the population. This coupled with efficient transmission (Hall, 2012; de Graaf et al., 2016) enabling infection of susceptible individuals likely therefore enables a virus to persist within a population for long time periods (Figure 4.5) and means that frequent inter-continental transmission is not necessary for persistence. Interestingly, we estimate that the rate of inter-continental transmission is highest prior to and in the first year of the pandemic and decreases throughout the pandemic (Figure 4.6). This is unlikely to be explained by a large number of cases within the first year of the pandemic resulting in more opportunities for migration due to the similarly high rate prior to the

pandemic, during a time at which the strain is present at low frequency. We therefore hypothesise that increasing herd immunity within the population during the pandemic may make it harder for a virus to migrate into the population.

We identified the substitutions that potentially enabled the emergence of the five most recent pandemic GII.4 strains by identifying sites that exhibited a nonsynonymous substitution leading to the strain common ancestor and a different amino acid residue(s) in the pandemic strain compared with the preceding pandemic strain (Figures 4.10-4.21). The emergence of a large number of pre-diversified lineages that likely account for a small proportion of all of the GII.4 lineages present at that point in time suggests that there is a phenotypic characteristic that is shared by and unique to these lineages. This is supported by the lack of evidence of particular lineages outcompeting other lineages within a pandemic strain (Figure 4.2), indicating each lineage has similar fitness within the human population. The large number of low level strains present at each point in time (Figure 3.1) suggests that the lineages that emerge have a phenotypic advantage, rather than each of the other low level strains having acquired a phenotypic disadvantage. Together, these observations support our hypothesis that the substitutions important for pandemic emergence were acquired by the root of the pandemic clade.

We identified three substitutions (S301N, A368E and N373H) that were potentially important for the pandemic emergence of Sydney 2012 (Figures 4.10, 4.11). Our results support previous experimental work on Sydney 2012 (Debbink et al., 2013) and further experiments are under way to validate the importance of these substitutions (as described in section 4.4.6). We could clearly identify the substitutions that occurred leading to the common ancestors of the New Orleans 2009 (Figure 4.12) and Farmington Hills 2002 (Figure 4.19) strains. These substitutions included changes at sites within and close to known epitopes in each case (Figures 4.12, 4.19). It has previously been suggested that sites close to known epitopes may alter viral antigenicity (Lindesmith et al., 2012a) and it is therefore possible that sites close to known epitopes may enable pandemic emergence. Identifying the substitutions that occurred leading to the Hunter 2004 and Den Haag 2006 pandemic strains was challenging due to uncertainty in the tree topology in each of these regions (Figures 2.1, 4.17). We were not able to conclusively determine the node that represents the common ancestor of Hunter 2004, although we did identify the nonsynonymous substitutions that occurred leading to each of the possible ancestors

(Figure 4.17). Accurate determination of the true Hunter 2004 ancestor will require experimental testing of the antigenic properties of the reconstructed sequences at nodes 1, 2 and 3 in Figure 4.17 and comparison with sampled viruses within the Hunter 2004 and Yerseke 2006 strains. We identified a large number of nonsynonymous substitutions that occurred leading to the Den Haag 2006 common ancestor (Figure 4.14), with these substitutions localising to multiple regions on the surface of the viral capsid (Figure 4.15). While we carried out homology modelling to predict the stuctural impact of substitutions leading to strain common ancestors (Figures 4.11, 4.16, 4.18, 4.20), we emphasise the importance of experimental validation to test the influence of each substitution on viral antigenicity. In particular, efforts are underway to reconstruct VLP panels for Farmington Hills 2002, Hunter 2004, Den Haag 2006 and New Orleans 2009 as described for Sydney 2012 in section 4.4.6. While we expect a change in antigenicity to be a major factor enabling pandemic emergence, substitutions may have alternative phenotypic effects that could promote pandemic spread, for example increasing particle stability or adaptability.

Interestingly, each of the pandemic strains exhibits a high degree of variability at the amino acid level (Figures 4.22, S4.4-S4.9 Table 4.3), with this variation commonly coinciding with putative epitopes (Figure 4.23) and HBGA-binding sites (Figure 4.24). New Orleans 2009 and Sydney 2012 began to accumulate diversity, including within epitopes and HBGA-binding sites, prior to pandemic emergence (Figures 4.25, 4.26) and the pandemic therefore involved the emergence of closely related, but genetically distinct, viruses. The onset of these pandemics therefore involved a soft selective sweep. We propose an extension to the three stage process of strain emergence where the low level pre-pandemic strain moves to a new region of antigenic space during the acquisition of changes phase (stage 1) due to one or more major changes (for example N310S, A368E and/or N373H in Sydney 2012) (Figure 4.28). It is not clear where in the antigenic space the pre-pandemic strain was prior to this point, but these important genetic changes result in movement to an antigenic region that is distinct to that of the current pandemic strain. The new region of antigenic space does not allow the pre-pandemic strain to emerge at that point in time, but rather the viruses in this region of the antigenic space will be able to emerge as a pandemic in the future. We hypothesise that emergence during this period is prevented by partial cross-reactivity of population immunity raised against the current pandemic strain and/or preceding pandemic/epidemic strains. During the diversification

phase, the strain acquires additional diversity that moves individual lineages within the local antigenic space, but each lineage remains in the same general region of antigenic space (Figure 4.28). At the onset of the pandemic, we hypothesise that all of the viruses within a certain region of the antigenic space can emerge, enabling the emergence of a diverse group of lineages that each share some antigenic characteristic(s). The region of antigenic space containing the viruses that can emerge will depend upon the immunity that has been built against the preceding pandemic strain, and potentially the pandemic and/or epidemic strains preceding that strain. Under this hypothesis, there are multiple regions of antigenic space that are occupied by low level viruses at any point in time and at the onset of the new pandemic, the strain with the best fit to the niche opened by the decline of the previous pandemic strain emerges.

The sites that exhibit high variability within a pandemic strain are typically specific to each strain, with only a small number of sites exhibiting high variability in more than one pandemic strain (Table 4.4). This indicates that this variation occurs either due to a strain-specific relaxation in selective constraints at the site or due to a strain-specific selective pressure to change at the site. While highly variable sites often coincide with epitope sites, the accumulation of diversity at such sites prior to pandemic emergence within New Orleans 2009 and Sydney 2012 suggests that variability is unlikely to be driven by a selection pressure to escape population immunity. Here, when the strain is present at low frequency there is likely a large susceptible population size and little population immunity to provide such a selection pressure. The variation within pandemic strains does not appear to correlate with geographical region as each of the common changes in Den Haag 2006, New Orleans 2009 and Sydney 2012 are found in viruses collected worldwide. Therefore the variation is unlikely to be driven by a selective pressure to better adapt to a specific geographical host population. However, site 393 is the only site that is variable in four of the six pandemic strains (Table 4.4) and is located within a loop required for binding to certain HBGAs (Singh et al., 2015). Indeed, substitutions at this site can alter HBGA binding (Lindesmith et al., 2008). Site 393 is highly variable between glycine and serine within Den Haag 2006, New Orleans 2009 and Sydney 2012 (Figures S4.7-S4.9). While the variation at this site does not correlate with geographical location, we hypothesise that viruses with glycine or serine at this site will have different HBGA-binding profiles. As individuals with the same HBGA expression are found worldwide,

substitutions at site 393 may provide a mechanism by which these strains can increase their susceptible population size and persist for longer.

4.6 Acknowledgements

The sequence of the Sydney 2012 reference VLP was provided by Lisa Lindesmith and Ralph Baric.

Chapter 5

The emerging GII.P16-GII.4 Sydney 2012 norovirus lineage is circulating worldwide, arose by late-2014 and contains polymerase changes that may increase virus transmission

5.1 Abstract

Noroviruses are the leading cause of human gastroenteritis worldwide. The norovirus genotype GII.4 causes most human outbreaks and has caused six pandemics of gastroenteritis since the mid-1990s. A novel norovirus lineage containing the pandemic GII.4 Sydney 2012 capsid in combination with the GII.P16 polymerase was recently detected in Asia and Germany. Here, we demonstrate that the GII.P16-GII.4 Sydney 2012 lineage is also circulating within the UK and USA and has been circulating since October 2014 or earlier. This lineage does not contain unique capsid substitutions but does contain substitutions in the polymerase that are close to positions known to influence polymerase function and virus transmission. These polymerase substitutions are shared with another novel norovirus lineage, GII.P16-GII.2, that also emerged as a major cause of norovirus outbreaks in Winter 2016-2017. Given the increase in prevalence of two novel lineages

that share polymerase substitutions but have different capsid genotypes, we suggest that these polymerase substitutions may have resulted in a more transmissible virus. Additionally we suggest that the combination of an advantageous polymerase with the pandemic GII.4 Sydney 2012 capsid may result in a highly transmissible virus. Further surveillance efforts will therefore be required to determine whether the GII.P16-GII.4 Sydney 2012 lineage increases in frequency and replaces the previously dominant GII.Pe-GII.4 Sydney 2012 virus over the coming months.

5.2 Introduction

The norovirus genotype GII.4 has been dominant within the human population since the mid-1990s, during which time it has caused six pandemics (Siebenga et al., 2007; van Beek et al., 2013). However, in the 2014-2015 Winter, a novel GII.17 lineage (termed Kawasaki 2014) emerged as the dominant cause of norovirus outbreaks in Asia (Matsushima et al., 2015; De Graaf et al., 2015; Lu et al., 2015). The origins of this lineage were traced back to Africa, with a single introduction into Asia (Lu et al., 2016). While sporadic cases of GII.17 Kawasaki 2014 have been detected from many countries worldwide since emergence in late 2014 (Chan et al., 2017a), this lineage has not replaced GII.4 Sydney 2012 as the dominant strain in most countries worldwide and therefore appears to be largely restricted to East Asia. More recently, two more novel norovirus lineages have emerged as major causes of norovirus outbreaks. The first of these lineages contains the GII.4 Sydney 2012 capsid with a GII.P16 RdRp, with reports of this lineage circulating within South Korea, Japan and Germany (Choi et al., 2017; Matsushima et al., 2016; Niendorf et al., 2017). While the GII.P16 RdRp is not typically prevalent, a GII.P16-GII.2 virus was the dominant strain amongst a large peak of norovirus cases and outbreaks in Germany in Winter 2016-2017 (Niendorf et al., 2017). Therefore these two novel lineages each contain the GII.P16 ORF1 but different capsid genotypes. The emergence of new GII.4 strains has often been associated with increased outbreaks, likely due to a lack of protective immunity against the new strain within the human population, raising concerns that the emergence of these novel lineages may result in increased outbreaks. Here, we demonstrate that the GII.P16-GII.4 Sydney 2012 lineage is also circulating within the

UK and USA. This lineage does not contain unique capsid substitutions, but does contain RdRp substitutions that are shared with the GII.P16-GII.2 RdRp and are close to positions known to influence RdRp function and viral transmission. These RdRp substitutions were acquired by March 2013. We conclude that these RdRp substitutions (possibly assisted by other substitutions in ORF1) have likely resulted in a more transmissible virus and have driven the increase in these novel norovirus lineages.

5.3 Materials and Methods

5.3.1 Sample collection, dataset assembly and sequencing

We identified noroviruses with the GII.P16 RdRp in ten stool samples collected as part of routine surveillance between June 2015 and April 2016. Samples were collected from South East and North West England from both sporadic cases and outbreaks. Four of these faecal specimens were referred to the Virus Reference Department, Public Health England, as part of a sentinel norovirus strain surveillance programme, which collects norovirus-positive specimens from geographically disparate regions across England. The other six faecal specimens were collected from a tertiary referral paediatric hospital in London, UK. RNA was extracted and whole genome sequencing performed as described previously (Brown et al., 2016b). For each sample, reads were mapped against a set of norovirus whole genome reference sequences spanning all known human genotypes to identify the best matching reference sequence. Whole genome sequences were assembled via reference mapping to this best matching reference. Consensus sequences were extracted from this mapping and used for all downstream analyses. Sample genotyping was carried out using the norovirus genotyping tool (Kroneman et al., 2011).

5.3.2 Phylogenetic analyses

Due to the presence of recombination close to ORF boundaries (Bull et al., 2007; Eden et al., 2013), we carried out all phylogenetic analyses on each ORF separately. We combined our sequences with all of the available sequences from the same genotype or strain available on GenBank with an available collection date as of 27/01/2017 to create datasets

for the GII.P16 ORF1 (n=45 including our samples), GII.P16 RdRp (n=179 including our samples), GII.4 Sydney 2012 capsid (n=717 including our samples) and GII.4 Sydney 2012 VP2 (n=176 including our samples). For samples without an accurate collection date, effort was made to identify the collection date with reference to the original literature. We aligned each dataset at the amino acid level using MUSCLE (Edgar, 2008) and reconstructed a nucleotide maximum likelihood tree for each dataset using RAxML (Stamatakis, 2014) with the GTR model of nucleotide substitution and gamma rate variation with 4 gamma classes. We assessed topological robustness using 1000 bootstrap replicates. Sequences that exhibit an excess of nucleotide change relative to their collection date may contain sequencing errors or a recombination breakpoint(s) or may have an incorrect collection date and the inclusion of such sequences can bias estimates of ancestral dates (Rambaut et al., 2016). Sequences that were overly divergent based on their collection date, as assessed using TempEst v1.5 (Rambaut et al., 2016), were removed from further analysis. We reconstructed the temporal evolutionary history in the GII.P16 RdRp and GII.4 Sydney 2012 capsid datasets using BEAST v2.4.2 (Bouckaert et al., 2014). For the GII.4 Sydney 2012 capsid dataset, we identified a well supported clade within the maximum likelihood tree (bootstrap support 81) that contains 70 sequences including all of the GII.P16-GII.4 Sydney 2012 sequences and ran BEAST on this smaller dataset. We used the most accurate collection date available for each sequence; the day of collection if available, the middle of the month of collection if the day of collection was not available or the middle of the year of collection if the month of collection was not available. We used the HKY model of nucleotide substitution with gamma rate heterogeneity with four rate classes. We used both the strict and relaxed lognormal clock models to test for variation in substitution rate across each clade. In each dataset, we employed a lognormal prior distribution on the substitution rate parameter. For the GII.4 Sydney 2012 capsid dataset, we set the mean of this prior distribution to 6.4×10^{-3}, which is the point estimate of the substitution rate parameter for the Sydney 2012 capsid estimated in chapter 3.4.3. For the GII.P16 RdRp dataset, we set the mean of the lognormal prior distribution to 3.1×10^{-3} which is the slope of the regression line between root-to-tip distance and sampling date calculated within TempEst v1.5 (Rambaut et al., 2016). In both cases, the standard deviation on the prior distribution was set to 0.1. We applied a coalescent Bayesian skyline tree prior. We carried out three replicate runs on each clock model with each dataset and ran

until convergence, as assessed using Tracer v1.6. The replicate runs were combined with removal of suitable burnin using LogCombiner v2.2.1 and the maximum clade credibility tree was identified using TreeAnnotator v2.2.1. To infer the nonsynonymous substitutions that occurred along branches of interest, we carried out ancestral reconstruction using the maximum likelihood tree reconstructed for each dataset. Initially, we used RAxML to optimise branch lengths in the maximum likelihood tree using the amino acid alignment with the WAG substitution model. Ancestral reconstruction was carried out using PAML v4.9 with the WAG substitution matrix and the nonsynonymous substitutions along branches of interest were inferred by comparing the reconstructed sequence at the start and end of the branch (Yang, 2007). The nonsynonymous substitutions that occurred leading to the novel GII.P16 clade in the RdRp are the same when using the ORF1 dataset and the RdRp dataset.

Ancestor dates were calculated using the complete posterior distribution of trees in each case. To identify the date at which the novel GII.P16 lineage diverged from the other sampled GII.P16 sequences, we calculated the date of the node immediately upstream of the common ancestor of the novel GII.P16 lineage in each tree in the posterior distribution and calculated the mean and 95% HPD of the distribution of this node date. To determine the date at which GII.4 Sydney 2012 acquired the GII.P16 RdRp, we identified the branch along which this acquisition occurred in each tree in the GII.4 Sydney 2012 capsid posterior distribution. We calculated the start date and end date of this branch and then calculated the mean and 95% HPD of this distribution of branch times. To identify whether the GII.P16-GII.3 sequences emerged from the GII.P16-GII.4 Sydney 2012 clade in the RdRp tree, we identified whether the GII.P16-GII.3 clade emerged from within the GII.P16-GII.4 Sydney 2012 clade in each tree in the posterior distribution. We defined the posterior probability as the proportion of trees that supported the GII.P16-GII.3 clade emerging from the GII.P16-GII.4 Sydney 2012 clade.

5.4 Results

5.4.1 The novel GII.P16 lineage has been circulating since March 2013 or earlier

Recent real time surveillance studies have identified the increased prevalence of viruses with the GII.P16 RdRp (Choi et al., 2017; Niendorf et al., 2017). We identified and whole genome sequenced ten viruses containing the GII.P16 RdRp that were collected in the UK as part of routine surveillance between June 2015 and April 2016. Genotyping demonstrated that seven of these viruses contained the GII.4 Sydney 2012 capsid and three contained the GII.3 capsid. The ten RdRp sequences cluster within a single clade in a RdRp maximum likelihood tree containing all available reference GII.P16 sequences (Figure 5.1). This clade also contains GII.P16-GII.4 Sydney 2012 sequences collected in the USA and Japan, including the GII.P16-GII.4 Sydney 2012 virus collected in 2016 in Kawasaki City, Japan (Matsushima et al., 2016). These RdRp sequences also cluster with the GII.P16-GII.2 RdRps collected in Germany in Winter 2016-2017 (Niendorf et al., 2017). There is a good correlation between root-to-tip distance and sampling date within the GII.P16 clade (Figure 5.1). We infer that the common ancestor of this GII.P16 clade occurred in March 2013 (95% HPD January 2012-May 2014, Figure 5.2), however this clade diverged from the other sampled GII.P16 sequences in January 2008 (95% HPD October 2004-March 2010). In both the maximum likelihood and Bayesian trees, it is well supported that the RdRps from the GII.P16-GII.3 and GII.P16-GII.4 Sydney 2012 sequences form a monophyletic cluster (Figure 5.1, 5.2). It is not clear, however, whether the GII.P16-GII.3 sequences emerged from within the GII.P16-GII.4 Sydney 2012 clade, with posterior support on this scenario of 0.44.

In a phylogenetic tree containing all GII.4 Sydney 2012 capsid sequences, the GII.P16-GII.4 Sydney 2012 sequences collected in the UK again cluster with sequences collected in the USA and Japan containing the GII.P16 RdRp (Figure 5.3). This clade contains additional capsid sequences collected in the USA where the RdRp was not sequenced. However, as all of the RdRps that have been sequenced within this clade are of the GII.P16 genotype, it is very likely that these samples also have the GII.P16 RdRp, although without additional sequencing and genotyping this remains speculative. There is

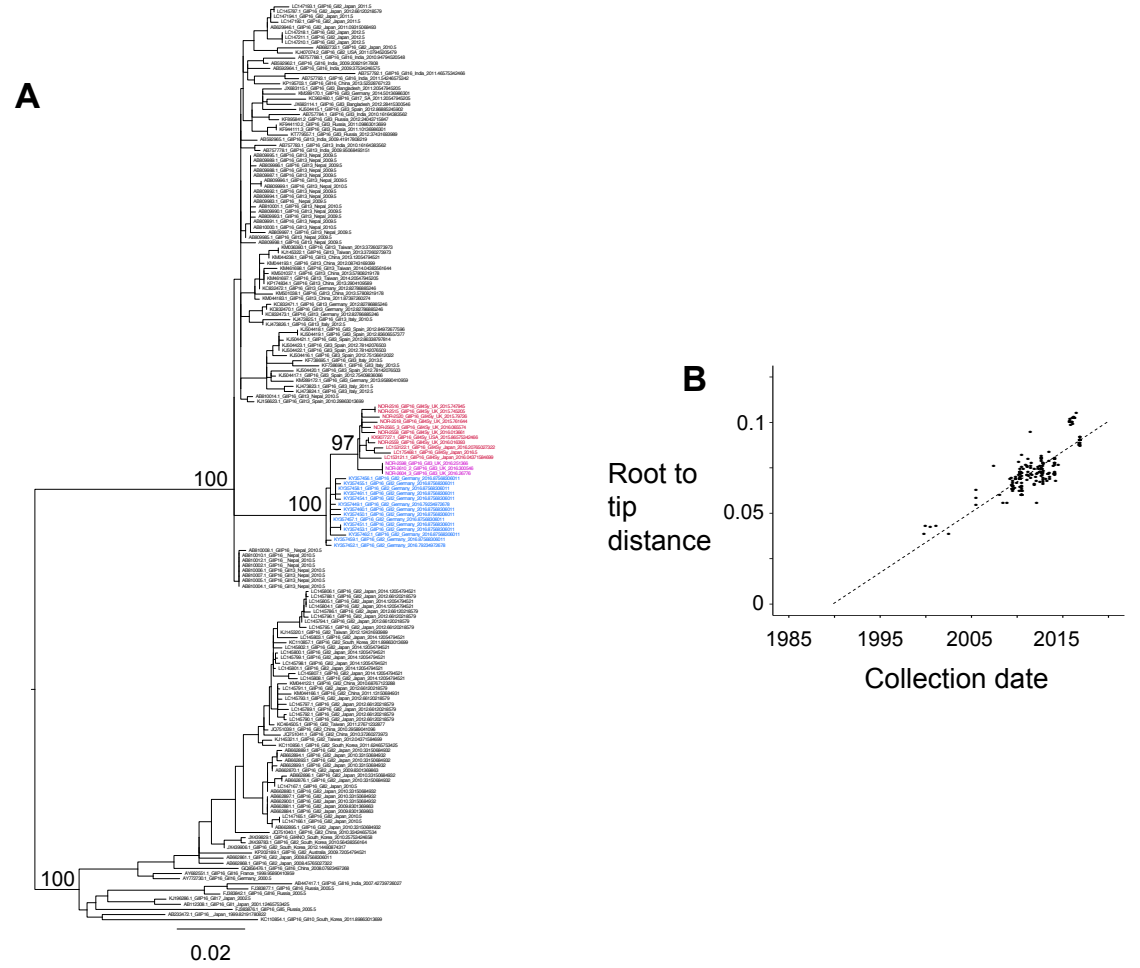


Figure 5.1: Maximum likelihood tree of the GII.P16 RdRp. **(A)** Nucleotide maximum likelihood tree of 179 GII.P16 RdRp sequences reconstructed using RAxML. Sequences within the novel GII.P16 lineage are coloured by capsid genotype: blue - GII.2, magenta - GII.3, red - GII.4 Sydney 2012. Bootstrap supports are shown at key nodes. **(B)** Correlation between root-to-tip distance and sampling date for the GII.P16 tree.

a correlation between the root-to-tip distance and sampling date within the GII.4 Sydney 2012 clade containing the GII.P16-GII.4 Sydney 2012 sequences (Figure 5.3). It is not clear whether GII.4 Sydney 2012 acquired the GII.P16 RdRp along the branch leading to the node in the tree where sequence KX354140.1 diverged (node A in Figure 5.3) or whether the GII.P16 RdRp was acquired after the divergence of sequence KX354140.1 (leading to node B in Figure 5.3) as the RdRp region was not sequenced in this sample. Given the long branch length leading to node A compared with the shorter branch length leading to node B, it is plausible that the GII.P16 RdRp was more likely acquired leading to node A. Node B occurred in October 2014 (95% HPD June 2014-February 2015) while node A occurred in May 2014 (95% HPD October 2013-October 2014). If the GII.P16

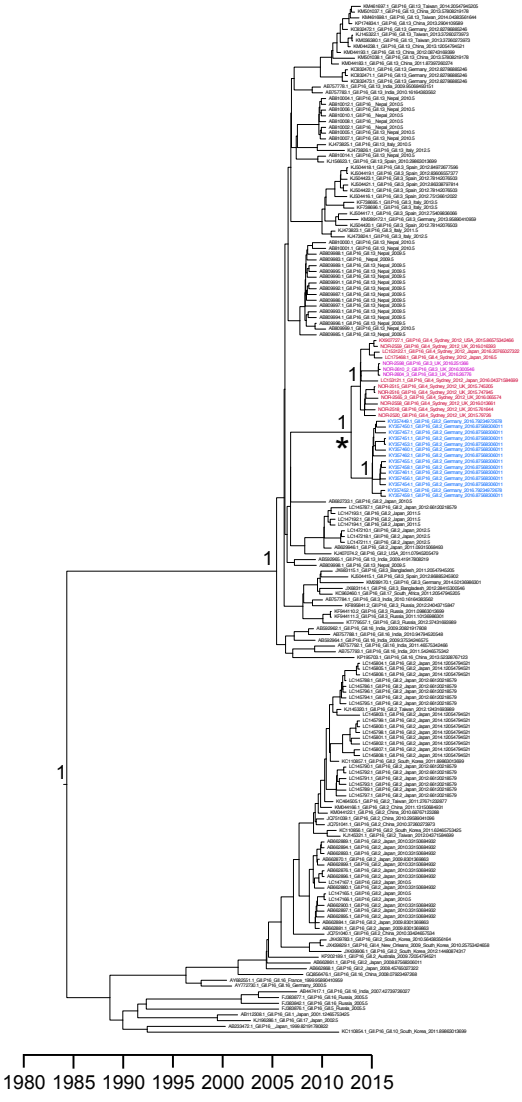


Figure 5.2: Temporal evolutionary history of the GII.P16 lineage. Time tree of 179 GII.P16 RdRp sequences reconstructed using BEAST 2. The sequences within the novel GII.P16 clade are coloured by the associated capsid genotype: blue - GII.2, magenta - GII.3, red - GII.4 Sydney 2012. The starred node is the common ancestor of this clade. Posterior supports are shown at key nodes.

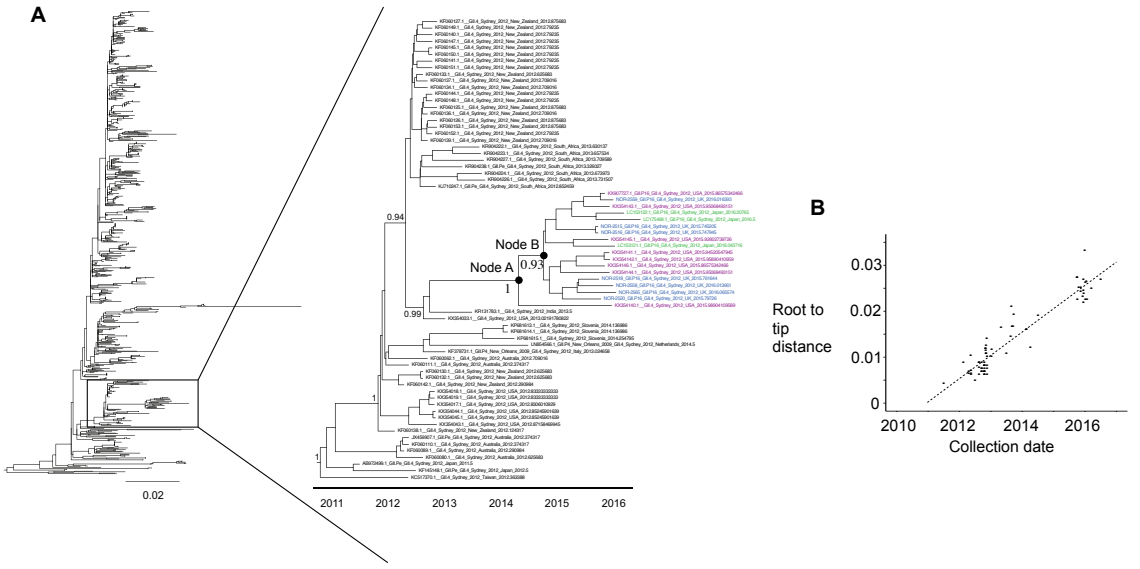


Figure 5.3: Evolutionary history of the GII.4 Sydney 2012 capsid. (A) A maximum likelihood phylogenetic tree was reconstructed for 781 GII.4 Sydney 2012 capsid sequences. From this, a well supported clade (bootstrap support 81) containing the GII.P16-GII.4 Sydney 2012 lineage and several other sequences was selected and a time tree reconstructed on this smaller dataset using BEAST 2. Node A and node B are the nodes that could be the common ancestor of the GII.P16-GII.4 Sydney 2012 clade; the RdRp region was not sequenced in sample KX354140.1 that branches from node A. The samples within the GII.P16-GII.4 Sydney 2012 lineage are coloured based on the country from which they were isolated: blue - UK, purple - USA, green - Japan. Posterior supports are shown at key nodes. **(B)** Correlation between root-to-tip distance and sampling date for the clade used in reconstruction of the time tree.

RdRp was acquired leading to node A, it was likely acquired in July 2013 (95% HPD September 2012-July 2014) while if the GII.P16 RdRp was acquired leading to node B, it was likely acquired in May 2014 (95% HPD February 2013-December 2014).

5.4.2 The novel GII.P16 lineage acquired substitutions in multiple ORF1 proteins, including the RdRp

No nonsynonymous substitutions occurred in the capsid leading to the GII.P16-GII.4 Sydney 2012 clade and there are no amino acid changes that are shared by the sequences in this clade that are not found in other Sydney 2012 capsids. Therefore the capsid sequences in this clade do not contain unique changes. However, we infer that 14 nonsynonymous substitutions occurred in ORF1 leading to the common ancestor of the GII.P16-GII.4 Sydney 2012/GII.3 clade (Figure 5.2, Table 5.1), with 11 of these changes being conserved amongst the sequences within this clade. Five of the sites at which substitutions occurred

are within the RdRp, with several of the sites occurring close to positions known to impact RdRp function and virus transmission (Figure 5.4) (Bull et al., 2010; Arias et al., 2016). Only the partial RdRp was sequenced from ORF1 in the GII.P16-GII.2 samples collected in Germany in Winter 2016-2017 (Niendorf et al., 2017) and this region only contains four of the sites that change leading to the common ancestor of the GII.P16-GII.4 Sydney 2012/GII.3 clade. However, all four of these nonsynonymous substitutions were also found within the GII.P16-GII.2 Germany sequences and these substitutions were therefore acquired by the common ancestor of the RdRp clade containing the GII.2, GII.3 and GII.4 Sydney 2012 sequences. We also infer a single nonsynonymous substitution (S157N) occurred in VP2 along the branch leading to the GII.P16-GII.4 Sydney 2012 clade.

Nonsynonymous change	Protein	RdRp position	Conserved within the novel GII.P16 clade
N52E	P48 (NS1/2)	N/A	Yes
S53P	P48 (NS1/2)	N/A	No
K165R	P48 (NS1/2)	N/A	Yes
S644P	NTPase (NS3)	N/A	No
R731K	P22 (NS4)	N/A	Yes
K750R	P22 (NS4)	N/A	Yes
P845Q	P22 (NS4)	N/A	Yes
A853T	P22 (NS4)	N/A	Yes
V1057I	Protease (NS6)	N/A	Yes
D1362E	RdRp (NS7)	173	Yes
S1482T	RdRp (NS7)	293	Yes
V1521I	RdRp (NS7)	332	No
K1546Q	RdRp (NS7)	357	Yes
T1549A	RdRp (NS7)	360	Yes

Table 5.1: Nonsynonymous substitutions in ORF1 leading to the novel GII.P16 clade. Each of the nonsynonymous substitutions that occurred leading to the common ancestor of the novel GII.P16 clade is shown with the nonstructural protein in which the site substitution occurred. Conserved sites have the same residue within each sequence in the novel GII.P16 clade.

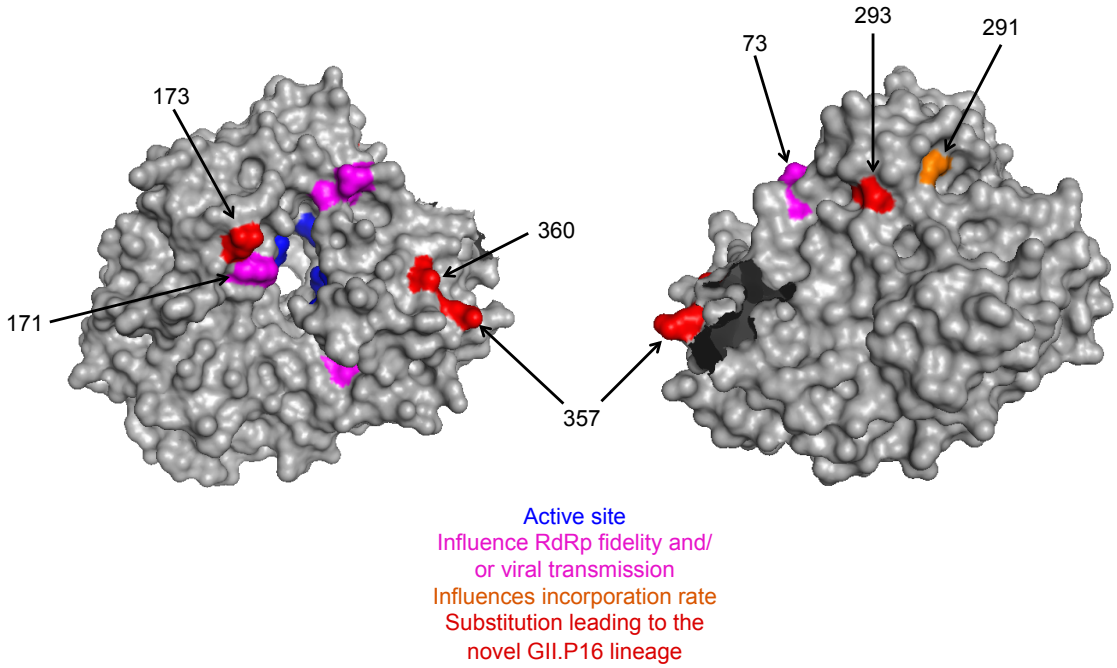


Figure 5.4: Location of RdRp sites that changed leading to the novel GII.P16 clade. The sites that changed leading to the novel GII.P16 lineage are shown in red. The sites that form the RdRp active site are shown in blue. Sites previously demonstrated to alter RdRp fidelity and/or viral transmission when mutated are shown in magenta (Arias et al., 2016). Site 291 was previously shown to alter the RdRp incorporation rate and is shown in orange (Bull et al., 2010). Sites are highlighted on PDB structure 1SH0, which is a structure of a GII.P4 RdRp (Ng et al., 2004).

5.5 Discussion

It has recently been reported that two novel norovirus lineages have become prevalent in outbreaks in several countries (Choi et al., 2017; Matsushima et al., 2016; Niendorf et al., 2017). Interestingly, these two novel lineages contain the GII.P16 ORF1 but different capsid genotypes. Here, we demonstrate that the GII.P16-GII.4 Sydney 2012 lineage is circulating within the UK and USA, in addition to the previous reports of circulation in Asia and Germany. We demonstrate, for the first time, that the same GII.P16 lineage is circulating in combination with the GII.3 capsid (Figure 5.1). This GII.P16 lineage is also found with the GII.2 capsid (Figure 5.1) and acquired substitutions in five of the six nonstructural proteins relative to other GII.P16 sequences (Table 1). Our phylogenetic analyses suggest that the novel GII.P16 lineage has been circulating since March 2013 or earlier and has been circulating with the GII.4 Sydney 2012 capsid since at least October 2014. Our analysis therefore suggests that the novel GII.P16 lineage circulated for at least

two years prior to its first identification. From the phylogenetic trees it is not clear which capsid genotype is likely to have been found ancestrally with the novel GII.P16 RdRp, with no clear support for the GII.P16 RdRps with one capsid genotype emerging from those with another capsid genotype (Figure 5.1, 5.2). However, the clustering in the tree does support each capsid genotype acquiring the GII.P16 RdRp through a single recombination event. The emergence of viruses with different capsid genotypes but a shared ORF1 lineage strongly suggests that it is the ORF1 that is important for the emergence of these strains and that one or more of the nonsynonymous changes that occurred leading to the common ancestor of the novel GII.P16 lineage may have resulted in a large increase in viral fitness (Table 5.1).

The increase in prevalence of the GII.P16 ORF1 with several different capsid genotypes is characteristic of a soft selective sweep driven by the GII.P16 ORF1. While little is known about the influence of several of the nonstructural proteins on viral transmission, previous studies have implicated the RdRp as an important component of viral fitness and it was recently demonstrated that changes in the RdRp can influence viral transmission by modulating replication fidelity and therefore viral diversity (Bull et al., 2010; Arias et al., 2016). Several of the substitutions leading to the common ancestor of the novel GII.P16 lineage are located within the palm domain of the RdRp and it is this region that contains most of the catalytic residues (Figure 5.4). While only four of the sites that changed leading to the novel GII.P16 lineage were sequenced in the GII.P16-GII.2 viruses from Germany, these viruses also contained the substitutions at these sites. We therefore suggest that the worldwide circulation of the GII.P16-GII.4 Sydney 2012 lineage and the high prevalence of viruses with the GII.P16 RdRp in the 2016-2017 Winter occurred due to increased transmissibility driven by RdRp changes in GII.P16. Testing the phenotypic consequences of the nonsynonymous substitutions acquired leading to the novel GII.P16 lineage in the recently developed human norovirus culture systems (Jones et al., 2014; Ettayebi et al., 2016) should be a priority.

The GII.2 capsid is typically rare and is therefore unlikely to be as fit as the prevalent GII.4 capsid. It is therefore possible that the combination of an advantageous RdRp with the GII.4 Sydney 2012 capsid may result in a highly transmissible virus. However, as the GII.P16-GII.4 Sydney 2012 lineage does not contain unique substitutions in the capsid region, this lineage is unlikely to be able to evade population immunity that has been

raised against the Sydney 2012 strain since its emergence as the pandemic strain in 2012. Whether the GII.P16-GII.4 Sydney 2012 lineage increases in prevalence in the future may therefore be determined by a balance between the benefit of combining an advantageous RdRp with a fit capsid and the decreased transmissibility of the Sydney 2012 capsid due to existing population immunity. Interestingly, since the time of this work GII.P16 viruses were found to also be prevalent in France, China and Taiwan during Winter 2016-2017 (Lu et al., 2017; Ao et al., 2017; Liu et al., 2017; Bidalot et al., 2017). There is a pattern emerging whereby the GII.P16-GII.2 strain is dominant in outbreaks amongst younger individuals (for example within kindergartens and schools), while GII.P16-GII.4 Sydney 2012 is dominant in outbreaks amongst older individuals (for example within nursing homes). This is consistent with the prevailing hypothesis that non-GII.4 capsid genotypes remain static through time and cannot therefore infect older individuals, while the GII.4 genotype has the capacity to evolve to evade host immunity and therefore can infect older individuals (Parra et al., 2017). The GII.P16-GII.3 viruses we detected were all isolated from children, which is also consistent with this hypothesis. This suggests that the GII.P16-GII.2 and GII.P16-GII.3 lineages are unlikely to be responsible for an increase in outbreaks in adults in the future. However, this further highlights the likely importance of the nonsynonymous substitutions in the GII.P16 ORF1, with the capsid appearing to act more as a vehicle to enable the virus to infect susceptible individuals.

Further surveillance will be required to determine whether the GII.P16-GII.2, GII.P16-GII.3 and GII.P16-GII.4 Sydney 2012 strains increase in frequency in the future. Importantly, surveillance strategies involving capsid genotyping alone will be unable to distinguish between the GII.P16-GII.4 Sydney 2012 capsid sequences and GII.4 Sydney 2012 capsid sequences circulating with the GII.P4 New Orleans 2009 or GII.Pe RdRps. It is therefore vital that surveillance efforts genotyping both the capsid and RdRp assess the prevalence of these strains over the coming months.

5.6 Acknowledgements

Stool samples were collected by David Allen at Public Health England and Julianne Brown at Great Ormond Street Hospital. cDNA preparation, sequencing library preparation and Illumina sequencing was carried out by the UCL Pathogen Genomics Unit.

Full genome assembly and consensus sequence extraction was carried out by Sunando Roy. All other analyses were carried out by me.

Chapter 6

Conclusions and future directions

The work in this thesis was concerned with understanding the evolution of GII.4 norovirus, the mechanism(s) through which this genotype circulates at high frequency within the human population and the sources and drivers of norovirus pandemics.

6.1 Understanding why the GII.4 genotype became pandemic in the mid-1990s

In chapter 2, we examined the factors that likely drove the increase in prevalence of the GII.4 capsid in the mid-1990s. While previous studies have suggested mechanisms that may enable GII.4 to be highly prevalent, this is the first study that has examined the non-synonymous substitutions that occurred leading to the pandemic GII.4 clade. Crucially, as these substitutions demarcate the low prevalence GII.4 lineages from the high prevalence GII.4 clade, they are likely responsible for this discrepancy in prevalence. Our results provided evidence that substitutions in the capsid and/or VP2 resulted in the increase in GII.4 prevalence. While we do hypothesise that the associated RdRp may be an important pre-requisite for high prevalence due to a putative high mutation rate, we demonstrated that this high mutation rate was likely acquired by 1906 and the RdRps found with pandemic GII.4 capsids last shared a common ancestor in the 1970s. Therefore while the associated RdRps may enable more efficient transmission, they did not drive the increase in prevalence in the mid-1990s. We identified the sites that were most likely responsible for the increase in RdRp substitution rate in 1906: 105 and 189. Future experiments to directly test the influence of F105Y, A189S and A189G mutations on the RdRp mutation rate and replication rate using previously developed assays (Bull et al., 2010) and/or in

the newly developed human norovirus culture systems (Jones et al., 2014; Ettayebi et al., 2016) would verify whether these changes truly drove the increase in substitution rate and which of the determinants of the substitution rate these RdRp substitutions affect.

The importance of substitutions in VP2 was hard to assess due to the paucity of information on VP2 functions and important residues for those functions. We identified sites 72, 78, 93, 148 and 187 in VP2 as being potentially important for the increase in GII.4 prevalence. In the future, the incorporation of the VP2 sequence at the common ancestor of the pandemic GII.4 clade and the sequence at the node upstream of this common ancestor into *in vitro* assays would enable probing of the functional relevance of changes at each of these sites. Given current evidence on the functions of VP2, it would be particularly interesting to test the stability of viral particles produced with each VP2 sequence and the efficiency of genome recruitment into viral particles.

While VP2 may have had a role in the increase in GII.4 frequency, we hypothesise that the major driver of this increase was substitutions in the capsid. We identified non-synonymous substitutions at 17 sites leading to the common ancestor of the pandemic GII.4 clade, although six of these sites exhibit the same amino acid residues in the pre-pandemic GII.4 lineages and the pandemic GII.4 clade and are therefore unlikely to be important. From these substitutions, we hypothesised two mechanisms by which the increase in frequency may have occurred: an increase in capsid stability and an increase in the range of HBGA-binding. The L333M and L459Q substitutions occurred at sites located in the dimer interface and we identified an increased interaction network formed by Q459 compared with S459 in solved crystal structures and inferred an increase in this network with Q459 compared with L459 in homology models. In the future, obtaining crystal structures of pre-pandemic GII.4 P domains with L at site 459 would confirm whether there was indeed an increase in the interaction network in this region leading to the pandemic GII.4 clade. Additionally, capsid sites 93, 172, 176, 295, 497 and 505 underwent a substitution leading to the pandemic GII.4 clade and then remained conserved throughout this clade. While these sites are not known to carry out particular functions, we hypothesised that these substitutions may have increased protein stability, providing a selective constraint not to change within the pandemic GII.4 clade. Such an increase in protein stability may have enabled the rapid accumulation of amino acid change we and others observe within the pandemic GII.4 clade. Our hypothesis of increased capsid

stability could be tested by measuring the length of time pre-pandemic GII.4 lineage and pandemic GII.4 clade viruses remain infectious upon surfaces or in groundwater. The newly developed cell culture systems (Jones et al., 2014; Ettayebi et al., 2016) would enable determination of whether the viruses remain replication-competent after set periods of time. Additionally, measuring the tolerance of sites throughout the capsid to mutation in the pandemic GII.4 clade and the pre-pandemic GII.4 lineages, either *in silico* or experimentally, would determine whether substitutions leading to the pandemic GII.4 clade enabled greater accumulation of amino acid change.

An increased HBGA-binding range has previously been hypothesised to enable the high prevalence of GII.4 viruses (Lindesmith et al., 2008; Donaldson et al., 2008), although here we provide a potential mechanism through the increased mobility of the 391-395 loop due to a decrease in residue size at site 395. Molecular dynamics work to examine the mobility of this loop with the pre-pandemic H residue at this site and the pandemic A, N and T residues at this site would be useful in the future. However, the high prevalence of GII.4 through to the present day suggests that the mechanisms responsible for high prevalence have been retained in each pandemic strain. Therefore the low efficiency of binding of Hunter 2004 to all tested HBGAs calls into question increased HBGA-binding as a mechanism of increased GII.4 prevalence. Future studies to determine firstly whether Hunter 2004 was less prevalent than other pandemic GII.4 strains and whether Hunter 2004 may have used alternate attachment factors are therefore necessary to determine whether increased HBGA-binding is truly a potential mechanism. While we identify these two mechanisms that may have resulted in an increase in prevlaence, there is clear potential for other mechanisms to have assisted in the increase.

In the final section in chapter 2, we compared the accumulation of nucleotide and amino acid change through time within GII.4 and other GII capsid genotypes. We found that while most GII genotypes accumulate change through time at the nucleotide level, GII.4 is the only genotype that accumulates change through time at the amino acid level. Concurrently with this work, another study was published with similar conclusions (Parra et al., 2017). However, we infer that the pre-pandemic GII.4 lineages were already accumulating amino acid change, albeit to a lower magnitude than the pandemic GII.4 clade. This suggests that the GII.4 capsid structure is inherently able to accomodate amino acid change in a way that the structures of the other GII capsid genotypes are not. This ability

to accomodate change may have been increased further leading to the pandemic GII.4 clade, as previously discussed. Confirmation of these results through estimation of nucleotide substitution rates and dN/dS ratios would be useful. Future studies examining exactly where within the capsid structure this amino acid change has been accumulated coupled with *in silico* analysis of the tolerance of capsid sequences from different genotypes to mutation may suggest a mechanism by which this change can be accomodated.

We conclude that the pandemic emergence of the GII.4 capsid genotype was a 'perfect storm' where a highly adaptable capsid structure found with a RdRp that enables rapid mutation and therefore efficient transmission acquired additional capsid stability and/or an increase in the susceptible population size through a broadened HBGA-binding range.

6.2 Understanding the sources and drivers of norovirus pandemics

In chapters 3 and 4, we examined the sources and drivers of the frequent norovirus pandemics that have occurred since the mid-1990s. Reconstruction of the temporal evolutionary history of the GII.4 genotype demonstrated that each of the pandemic and epidemic GII.4 strains was present and hidden for years prior to causing their respective pandemic or epidemic. This provided evidence for the first time that none of the pandemic strains evolved directly from a preceding pandemic strain. Rather, pandemic strains diverge from other strains years prior to emergence and evolve independently in one or more currently poorly sampled reservoir populations. This was supported by the identification of pre-pandemic sequences from multiple pandemic strains. We demonstrated that each of the pandemic strains diverged into a large number of lineages over several years prior to pandemic emergence and that the onset of the new pandemic is therefore caused by the concurrent emergence of a large number of closely related viruses. This indicates that the pandemic-enabling genetic changes are acquired years prior to pandemic spread and are therefore not the proximal driver of the pandemic. We instead hypothesised that a new pandemic is driven by changes in host immunity that enable the emergence of a pre-adapted GII.4 strain (Figure 6.1). We finished this chapter by proposing a three stage process of GII.4 pandemic strain emergence where 1) substitutions are acquired that will

enable pandemic emergence in the future, 2) the strain diversifies into a large number of lineages containing the pre-adaptation changes and 3) the multiple closely related lineages emerge to cause the new pandemic. Our results suggest that increased surveillance efforts of the community and poorly sampled geographical regions are required and coupling such surveillance with methods to compare viral antigenicity has the potential to detect strains that have the potential to become pandemic in the future before they actually do. These results additionally demonstrate the importance of a broadly targeting vaccine against GII.4 norovirus as a vaccine targeted against an individual strain may increase host herd immunity and hasten the onset of the next pandemic.

In chapter 4, we addressed questions raised by our analyses in chapter 3, in particular examining potential source regions of norovirus pandemics and the viral genetic changes that may enable a new pandemic to occur. We carried out a phylogeographic analysis of the New Orleans 2009 and Sydney 2012 strains and demonstrated wide and consistent circulation prior to pandemic emergence. In particular, each strain was introduced into each continent prior to pandemic emergence and in most cases was first imported at least two years prior to the pandemic. After introduction, the strain continued circulating within the continent until at least the date of the latest sequence in out dataset from that continent. Therefore these strains underwent consistent low-level worldwide circulation for several years prior to emerging pandemically. We found no evidence of a consistent source region for the origin or early circulation of pandemic GII.4 strains but did identify Asia as a net source of viral lineages in each strain. Our phylogeographic analyses also identified that viral lineages typically persist within a continent for roughly two years. Therefore we suggest that the large peak of outbreaks within each winter season in temperate countries is mostly caused by viruses that have persisted within the region since the previous winter.

We carried out an assessment of the nonsynonymous substitutions that occurred leading to the common ancestor of the five most recent pandemic GII.4 strains. These substitutions are very likely to have enabled the pandemic emergence of the strain as they represent the changes that are specific to that clade of viruses compared with the other low level pre-pandemic strains that our analysis in chapter 3 suggested are also present through time. We hypothesised mechanisms that may have enabled the pandemic emergence of each strain by examining homology models. Future studies employing monoclonal antibodies and polyclonal sera that have been raised against each pandemic strain will be

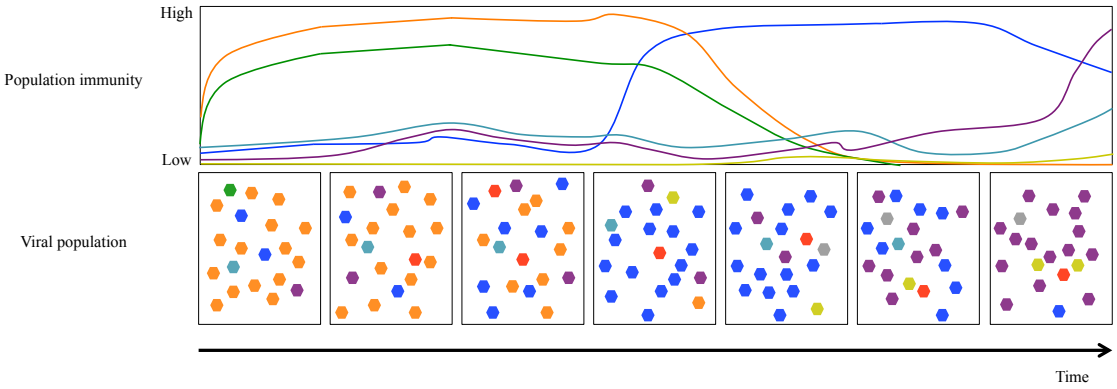


Figure 6.1: Changes in strain frequency and population immunity through time. Our results in chapter 3 suggested that there are a large number of low level distinct GII.4 strains present through time. Here, different coloured hexagons represent antigenically distinct GII.4 strains. The different pairs of strains have more or less cross-reactivity of immunity between them depending on the pair, for example here the orange and green strains are similar antigenically and so immunity against one strain is highly cross-reactive against the other strain, while there is little cross-reactivity between the orange and blue strains. At the start of the figure, the orange strain has recently become pandemic and there is therefore rapidly increasing population immunity against this strain. Due to the high cross-reactivity of immunity between the orange and green strains, there is also a rapid increase in immunity against the green strain at this point in time, even though the green strain is only present at low frequency in the population. Due to little cross-reactivity between the orange and blue strains, the immunity against the blue strain remains low. When population immunity against the orange strain is high, this curtails the spread of the orange strain, thereby opening a niche into which another strain can emerge. We hypothesise that the strain that emerges, in this case the blue strain, is the strain that combines the best opportunity to spread with the least cross-reactivity of immunity against the preceding pandemic strain, and potentially against the pandemic and epidemic strains preceding that. Immunity against the blue strain then increases in the population as the blue strain is highly prevalent. However, as norovirus immunity is likely to be relatively short lasting, the immunity against the orange and green strains decreases through time. As the immunity against the blue strain is high, the spread of the blue strain is curtailed, opening the opportunity for the purple strain, which combines the best opportunity to spread with the least cross-reactivity of immunity against the blue strain and other previous strains, to emerge.

useful to test which of the substitutions we identified alter viral antigenicity and may have therefore enabled pandemic emergence. Such experiments testing the influence of sites 310, 368 and 373 in the pandemic emergence of Sydney 2012 are already underway. We identified significant amino acid variation within each of the pandemic GII.4 strains, with highly variable sites often coinciding with epitope regions and/or HBGA-binding sites. It is therefore possible that this variation alters important viral phenotypes, including antigenicity and the susceptible population size. Importantly, diversity, including at antigenic and HBGA-binding sites, began to be accumulated in New Orleans 2009 and Sydney 2012 prior to pandemic emergence. Future experiments measuring antigeniticy and HBGA-binding profiles would confirm this; again, these studies are underway for Sydney 2012.

We incorporated the phylogeographic and analysis of amino acid substitution results into our three stage process of strain emergence. During the diversification phase, the lineages undergo spread to multiple geographically distinct reservoirs from which they emerge at the onset of the pandemic. We hypothesise that one or more of the substitutions leading to the pandemic strain common ancestor results in a large change in antigenicity and moves the virus to a new region of antigenic space. The strain then diversifies resulting in viruses with subtly different antigenicity but these substitutions do not move the virus out of the vicinity of the new region of antigenic space. At the onset of the new pandemic, all of the viruses within the new region of antigenic space emerge to cause the pandemic. The concurrent emergence of multiple closely related lineages that have been circulating on different continents adds support to our hypothesis that changes in host immunity drive new pandemics.

In summary, our results in chapter 3 demonstrated that there are a large number of low level pre-pandemic and pre-epidemic GII.4 strains that are present through time (Figure 6.1). Each of these strains is in the process of diversifying into multiple lineages, with each of these lineages either persisting within the population for a period of time or becoming extinct. However, we hypothesise that these low level strains are prevented from gaining a foothold in the population due to partial cross reactivity of immunity, which is sufficient to prevent a rare strain from gaining a foothold. Once sufficient immunity has been built against the current pandemic strain, this strain can no longer spread efficiently and so begins to decline in prevalence. This decline in prevalence results in a decline in

immunity against the current pandemic which enables the next pandemic strain to gain a foothold in the population and increase in prevalence to become the dominant strain in outbreaks. The low level strain that emerges at that point in time is the one that combines the best opportunity to emerge with the least cross-reactivity against pre-existing immunity. All of the pre-diversified lineages within the advanatageous region of the antigenic space can emerge at this point in time. Importantly, due to the universally high prevalence of GII.4 norovirus, population immunity can build across the world at a relatively constant rate. In the future, we plan to test this hypothesis using a combination of viral sequences, phylogenetics, human monoclonal antibodies and polyclonal sera raised against individual GII.4 pandemic strains, antigenic cartography and mathematical modelling. We will measure the strength of binding of population sera collected through time against a set of sequences from each of the pandemic GII.4 strains. This will determine whether there are antigenic shifts between pandemic GII.4 strains, the magnitude of such shifts and how much antigenic variation there is within a pandemic GII.4 strain (Figure 6.2). We will also test the antigenic properties of ancestral viruses from each strain. We would expect to find a large antigenic difference between each of the pandemic strains, as per the shift in the region of antigenic space in our hypothesis of strain emergence (Figure 6.2). We hypothesise that the strain common ancestor, downstream ancestors and sampled tip sequences will form a cluster within the antigenic space that is distinct from sequences in other pandemic strains. As we have demonstrated that the strain common ancestor and downstream ancestors occurred prior to the onset of the pandemic, this would provide support that important genetic changes for strain emergence are acquired years prior to the pandemic and for the local movement within antigenic space during the diversification phase of strain emergence. A hypothetical example including the Den Haag 2006, New Orleans 2009 and Sydney 2012 strains is shown in figure 6.2. We will then incorporate these results into a mathematical framework that will incorporate changes in host immunity through time to test the hypothesis that changes in host immunity can drive the emergence of new pandemic strains and the circumstances under which that could occur.

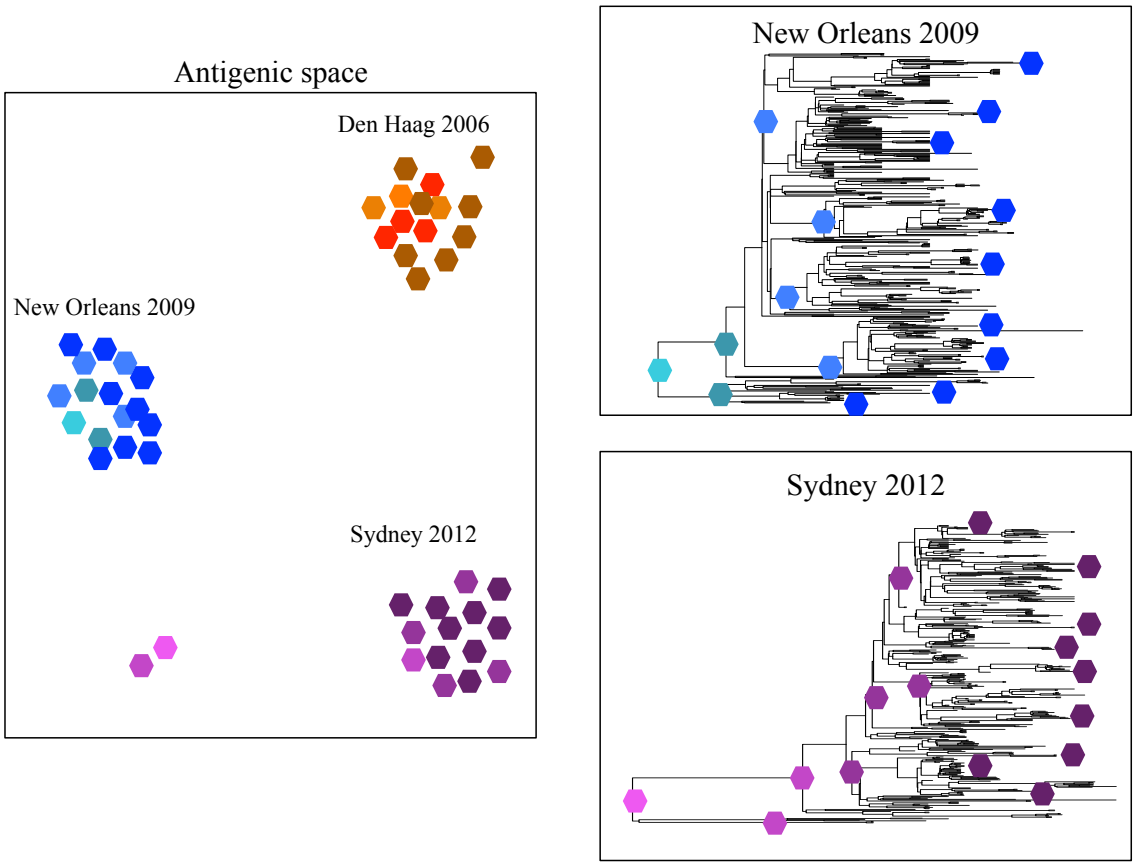


Figure 6.2: Testing the antigenic properties of pandemic GII.4 strains. In the future, we will test the antigenic relationship between sequences and reconstructed ancestor sequences from each of the pandemic strains using population polyclonal sera collected through time. We will measure the binding affinity of each serum sample against each viral sequence and map this binding data into an antigenic space using multidimensional scaling. A hypothetical scenario is shown here for Den Haag 2006, New Orleans 2009 and Sydney 2012. Under our hypothesis, the viruses within a strain are antigenically similar and this shared antigenic characteristic(s) is acquired by the root of the pandemic clade, while the different pandemic GII.4 strains are antigenically distinct. We therefore expect the reconstructed ancestor sequences to cluster with the tip sequences in antigenic space. We suggested in chapter 4 that an antigenic change occurred between the common ancestor of all Sydney 2012 sequences and the common ancestor of the pandemic Sydney 2012 clade. Therefore the common ancestor of all Sydney 2012 sequences and one of the sequences immediately downstream of this ancestor are shown in a distinct region of antigenic space to the remainder of the Sydney 2012 sequences. Such a set of results would provide support for the three stage process of strain emergence described in chapters 3 and 4. An example set of reconstructed ancestral sequences and tip sequences whose antigenic properties we would test is shown on the New Orleans 2009 and Sydney 2012 phylogenetic trees.

6.3 The pandemic potential of viral lineages with the GII.P16 ORF1

In chapter 5 we examined two norovirus strains that emerged as major causes of gastroenteritis in Winter 2016-2017. These strains each contained the GII.P16 ORF1 but contained different capsid genotypes; GII.2 and GII.4 Sydney 2012. We demonstrated that the GII.P16-GII.4 Sydney 2012 lineage is circulating in the UK and USA, in addition to the previously identified circulation in Asia and Germany. We showed that the GII.P16 RdRps found within the two novel lineages form a monophyletic clade that also includes GII.P16-GII.3 sequences from the UK and coalesces to a common ancestor in March 2013. There were no unique substitutions within the Sydney 2012 capsid in this lineage, suggesting this strain will not be able to evade existing herd immunity raised against earlier circulating Sydney 2012 viruses. We did, however, identify substitutions within ORF1 that are shared between the several strains with the novel GII.P16 ORF1. Importantly, these substitutions include sites within the RdRp that are close to sites known to influence RdRp function and viral transmission and so may have resulted in a highly transmissible virus. In the future, testing the effect of these substitutions on viral replication in the human norovirus culture systems (Jones et al., 2014; Ettayebi et al., 2016) would determine the functional importance of these substitutions. Additionally, it is vital that surveillance targeting ORF1 and ORF2 is carried out to determine whether these novel GII.P16 lineages increase in prevalence.

Bibliography

Ahmed, S. M., Hall, A. J., Robinson, A. E., Verhoef, L., Premkumar, P., Parashar, U. D., Koopmans, M., and Lopman, B. A. (2014). Global prevalence of norovirus in cases of gastroenteritis: A systematic review and meta-analysis. *The Lancet Infectious Diseases*, 14(8):725–730.

Ahmed, S. M., Lopman, B. A., and Levy, K. (2013). A Systematic Review and Meta-Analysis of the Global Seasonality of Norovirus. *PLoS ONE*, 8(10).

Alhatlani, B., Vashist, S., and Goodfellow, I. (2015). Functions of the 5' and 3' ends of calicivirus genomes. *Virus Research*, 206:134–143.

Allen, D. J., Adams, N. L., Aladin, F., Harris, J. P., and Brown, D. W. G. (2014). Emergence of the GII-4 norovirus Sydney2012 strain in England, winter 2012-2013. *PLoS ONE*, 9(2).

Allen, D. J., Gray, J. J., Gallimore, C. I., Xerry, J., and Iturriza-Gómara, M. (2008). Analysis of amino acid variation in the P2 domain of the GII-4 Norovirus VP1 protein reveals putative variant-specific epitopes. *PLoS ONE*, 3(1).

Allen, D. J., Trainor, E., Callaghan, A., O'Brien, S. J., Cunliffe, N. A., and Iturriza-G??mara, M. (2016). Early detection of epidemic GII-4 norovirus strains in UK and Malawi: Role of surveillance of sporadic acute gastroenteritis in anticipating global epidemics. *PLoS ONE*, 11(4).

Ando, T., Noel, J. S., and Fankhauser, R. L. (2000). Genetic classification of "Norwalk-like viruses. *The Journal of infectious diseases*, 181 Suppl(Suppl 2):S336–S348.

Ao, Y., Wang, J., Ling, H., He, Y., Dong, X., Wang, X., Peng, J., Zhang, H., Jin, M.,

and Duan, Z. (2017). Norovirus GII.P16/GII.2Associated Gastroenteritis, China, 2016. *Emerging Infectious Diseases*, 23(7):1172–1175.

Arias, A., Thorne, L., Ghurburrun, E., Bailey, D., and Goodfellow, I. (2016). Norovirus Polymerase Fidelity Contributes to Viral Transmission In Vivo. *mSphere*, 1(5).

Atmar, R. L., Bernstein, D. I., Harro, C. D., Al-Ibrahim, M. S., Chen, W. H., Ferreira, J., Estes, M. K., Graham, D. Y., Opekun, A. R., Richardson, C., and Mendelman, P. M. (2011). Norovirus Vaccine against Experimental Human Norwalk Virus Illness. *New England Journal of Medicine*, 365(23):2178–2187.

Atmar, R. L., Bernstein, D. I., Lyon, G. M., Treanor, J. J., Al-Ibrahim, M. S., Graham, D. Y., Vinjé, J., Jiang, X., Gregoricus, N., Frenck, R. W., Moe, C. L., Chen, W. H., Ferreira, J., Barrett, J., Opekun, A. R., Estes, M. K., Borkowski, A., Baehner, F., Goodwin, R., Edmonds, A., and Mendelman, P. M. (2015). Serological correlates of protection against a GII.4 norovirus. *Clinical and Vaccine Immunology*, 22(8):923–929.

Atmar, R. L., Opekun, A. R., Gilger, M. A., Estes, M. K., Crawford, S. E., Neill, F. H., and Graham, D. Y. (2008). Norwalk virus shedding after experimental human infection. *Emerging Infectious Diseases*, 14(10):1553–1557.

Atmar, R. L., Opekun, A. R., Gilger, M. A., Estes, M. K., Crawford, S. E., Neill, F. H., Ramani, S., Hill, H., Ferreira, J., and Graham, D. Y. (2014). Determination of the 50% human infectious dose for norwalk virus. *Journal of Infectious Diseases*, 209(7):1016–1022.

Bailey, D., Kaiser, W. J., Hollinshead, M., Moffat, K., Chaudhry, Y., Wileman, T., Sosnovtsev, S. V., and Goodfellow, I. G. (2010). Feline calicivirus p32, p39 and p30 proteins localize to the endoplasmic reticulum to initiate replication complex formation. *Journal of General Virology*, 91(3):739–749.

Baric, R., Yount, B., and Lindesmith, L. (2002). Expression and self-assembly of Norwalk virus capsid protein from Venezuelan equine encephalitis virus replicons. *Journal of Virology*, 76(6):3023 – 3030.

Bartnicki, E., Cunha, J. B., Kolawole, A. O., and Wobus, C. E. (2017). Recent advances in understanding noroviruses. *F1000Research*, 6:79.

Bartsch, S. M., Lopman, B. A., Hall, A. J., Parashar, U. D., and Lee, B. Y. (2012). The potential economic value of a human norovirus vaccine for the United States. *Vaccine*, 30(49):7097–7104.

Bartsch, S. M., Lopman, B. A., Ozawa, S., Hall, A. J., and Lee, B. Y. (2016). Global economic burden of norovirus gastroenteritis. *PLoS ONE*, 11(4).

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D. J., Suchard, M. A., Tashiro, M., Wang, D., Xu, X., Lemey, P., and Russell, C. A. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220.

Belliot, G., Sosnovtsev, S. V., Chang, K.-O., Babu, V., Uche, U., Arnold, J. J., Cameron, C. E., and Green, K. Y. (2005). Norovirus proteinase-polymerase and polymerase are both active forms of RNA-dependent RNA polymerase. *Journal of virology*, 79(4):2393–2403.

Belliot, G., Sosnovtsev, S. V., Mitra, T., Hammer, C., Garfield, M., and Green, K. Y. (2003). In vitro proteolytic processing of the MD145 norovirus ORF1 nonstructural polyprotein yields stable precursors and products similar to those detected in calicivirus-infected cells. *Journal of virology*, 77(20):10957–10974.

Bernstein, D. I., Atmar, R. L., Lyon, G. M., Treanor, J. J., Chen, W. H., Jiang, X., Vinjé, J., Gregoricus, N., Frenck, R. W., Moe, C. L., Al-Ibrahim, M. S., Barrett, J., Ferreira, J., Estes, M. K., Graham, D. Y., Goodwin, R., Borkowski, A., Clemens, R., and Mendelman, P. M. (2015). Norovirus vaccine against experimental human GII.4 virus illness: A challenge study in healthy adults. *Journal of Infectious Diseases*, 211(6):870–878.

Bertolotti-Ciarlet, A., Crawford, S. E., Hutson, A. M., and Estes, M. K. (2003). The 3' end of Norwalk virus mRNA contains determinants that regulate the expression and stability of the viral capsid protein VP1: a novel function for the VP2 protein. *Journal of virology*, 77(21):11603–11615.

Bertolotti-Ciarlet, A., White, L. J., Chen, R., Prasad, V., Estes, M. K., and Prasad, B. V. V. (2002). Structural Requirements for the Assembly of Norwalk Virus-Like Particles. *Journal of Virology*, 76(8):4044–4055.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., and Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1).

Bidalot, M., Théry, L., Kaplon, J., De Rougemont, A., and Ambert-Balay, K. (2017). Emergence of new recombinant noroviruses GII.P16-GII.4 and GII.p16-GII.2, France, winter 2016 to 2017. *Eurosurveillance*, 22(15).

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874.

Bok, K., Abente, E. J., Realpe-Quintero, M., Mitra, T., Sosnovtsev, S. V., Kapikian, A. Z., and Green, K. Y. (2009). Evolutionary dynamics of GII.4 noroviruses over a 34-year period. *Journal of virology*, 83(22):11890–901.

Bok, K. and Green, K. (2012). Norovirus gastroenteritis in immunocompromised patients. *New England Journal of Medicine*, 119(8):1831–1837.

Bok, K., Parra, G. I., Mitra, T., Abente, E., Shaver, C. K., Boon, D., Engle, R., Yu, C., Kapikian, A. Z., Sosnovtsev, S. V., Purcell, R. H., and Green, K. Y. (2011). Chimpanzees as an animal model for human norovirus infection and vaccine development. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):325–330.

Boon, D., Mahar, J. E., Abente, E. J., Kirkwood, C. D., Purcell, R. H., Kapikian, A. Z., Green, K. Y., and Bok, K. (2011). Comparative Evolution of GII.3 and GII.4 Norovirus over a 31-Year Period. *Journal of Virology*, 85(17):8656–8666.

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2008). Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, 4(1):1–13.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, 10(4).

Brown, J. R., Gilmour, K., and Breuer, J. (2016a). Norovirus Infections Occur in B-Cell-Deficient Patients. *Clinical Infectious Diseases*, 62(9):1136–1138.

Brown, J. R., Roy, S., Ruis, C., Yara Romero, E., Shah, D., Williams, R., and Breuer, J. (2016b). Norovirus whole-genome sequencing by SureSelect target enrichment: A robust and sensitive method. *Journal of Clinical Microbiology*, 54(10):2530–2537.

Brown, J. R., Shah, D., and Breuer, J. (2016c). Viral gastrointestinal infections and norovirus genotypes in a paediatric UK hospital, 20142015. *Journal of Clinical Virology*, 84:1–6.

Bu, W., Mamedova, A., Tan, M., Xia, M., Jiang, X., and Hegde, R. S. (2008). Structural basis for the receptor binding specificity of Norwalk virus. *Journal of virology*, 82(11):5340–7.

Bucardo, F., Kindberg, E., Paniagua, M., Vildevall, M., and Svensson, L. (2009). Genetic susceptibility to symptomatic norovirus infection in nicaragua. *Journal of Medical Virology*, 81(4):728–735.

Bull, R. a., Eden, J.-S., Luciani, F., McElroy, K., Rawlinson, W. D., and White, P. a. (2012). Contribution of Intra- and Interhost Dynamics to Norovirus Evolution. *Journal of Virology*, 86(6):3219–3229.

Bull, R. A., Eden, J. S., Rawlinson, W. D., and White, P. A. (2010). Rapid evolution of pandemic noroviruses of the GII.4 lineage. *PLoS Pathogens*, 6(3).

Bull, R. A., Hansman, G. S., Clancy, L. E., Tanaka, M. M., Rawlinson, W. D., and White, P. A. (2005). Norovirus recombination in ORF1/ORF2 overlap. *Emerging Infectious Diseases*, 11(7):1079–1085.

Bull, R. A., Tanaka, M. M., and White, P. A. (2007). Norovirus recombination. *Journal of General Virology*, 88(12):3347–3359.

Bull, R. A., Tu, E. T. V., McIver, C. J., Rawlinson, W. D., and White, P. A. (2006). Emergence of a new norovirus genotype II.4 variant associated with global outbreaks of gastroenteritis. *Journal of Clinical Microbiology*, 44(2):327–333.

Bull, R. A. and White, P. A. (2011). Mechanisms of GII.4 norovirus evolution.

Caddy, S. L., de Rougemont, A., Emmott, E., El-Attar, L., Mitchell, J. A., Hollinshead, M., Belliot, G., Brownlie, J., Le Pendu, J., and Goodfellow, I. (2015). Evidence for human norovirus infection of dogs in the UK. *Journal of Clinical Microbiology*, 53(April):JCM.02778–14.

Cannon, J. L., Lindesmith, L. C., Donaldson, E. F., Saxe, L., Baric, R. S., and Vinjé, J. (2009). Herd immunity to GII.4 noroviruses is supported by outbreak patient sera. *Journal of virology*, 83(11):5363–74.

Cao, S., Lou, Z., Tan, M., Chen, Y., Liu, Y., Zhang, Z., Zhang, X. C., Jiang, X., Li, X., and Rao, Z. (2007). Structural basis for the recognition of blood group trisaccharides by norovirus. *Journal of virology*, 81(11):5949–5957.

Carlsson, B., Kindberg, E., Buesa, J., Rydell, G. E., Lidón, M. F., Montava, R., Mallouh, R. A., Grahn, A., Rodríguez-Díaz, J., Bellido, J., Arnedo, A., Larson, G., and Svensson, L. (2009). The G428A nonsense mutation in FUT2 provides strong but not absolute protection against symptomatic GII.4 norovirus infection. *PLoS ONE*, 4(5).

Chachu, K. A., LoBue, A. D., Strong, D. W., Baric, R. S., and Virgin, H. W. (2008a). Immune mechanisms responsible for vaccination against and clearance of mucosal and lymphatic norovirus infection. *PLoS Pathogens*, 4(12).

Chachu, K. A., Strong, D. W., LoBue, A. D., Wobus, C. E., Baric, R. S., and Virgin, H. W. (2008b). Antibody Is Critical for the Clearance of Murine Norovirus Infection. *Journal of Virology*, 82(13):6610–6617.

Chan, M. C., Hu, Y., Chen, H., Podkolzin, A. T., Zaytseva, E. V., Komano, J., Sakon, N., Poovorawan, Y., Vongpunsawad, S., Thanusuwannasak, T., Hewitt, J., Croucher, D., Collins, N., Vinjé, J., Pang, X. L., Lee, B. E., de Graaf, M., van Beek, J., Vennema, H., Koopmans, M. P., Niendorf, S., Poljsak-Prijatelj, M., Steyer, A., White, P. A., Lun,

J. H., Mans, J., Hung, T.-N., Kwok, K., Cheung, K., Lee, N., and Chan, P. K. (2017a). Global Spread of Norovirus GII.17 Kawasaki 308, 20142016. *Emerging Infectious Diseases*, 23(8):1359–1354.

Chan, M. C. W., Kwok, K., Hung, T.-N., Chan, L.-Y., and Chan, P. K. S. (2017b). Complete Genome Sequence of an Emergent Recombinant GII.P16-GII.2 Norovirus Strain Associated with an Epidemic Spread in the Winter of 2016- 2017 in Hong Kong, China. *Genome Announcements*, 5(20):16–17.

Chan, M. C. W., Lee, N., Hung, T.-N., Kwok, K., Cheung, K., Tin, E. K. Y., Lai, R. W. M., Nelson, E. A. S., Leung, T. F., and Chan, P. K. S. (2015). Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII.17 variant in late 2014. *Nature Communications*, 6:10061.

Chaudhry, Y., Nayak, A., Bordeleau, M. E., Tanaka, J., Pelletier, J., Belsham, G. J., Roberts, L. O., and Goodfellow, I. G. (2006). Caliciviruses differ in their functional requirements for eIF4F components. *Journal of Biological Chemistry*, 281(35):25315–25325.

Cheesbrough, J. S., Green, J., Gallimore, C. I., Wright, P. a., and Brown, D. W. (2000). Widespread environmental contamination with Norwalk-like viruses (NLV) detected in a prolonged hotel outbreak of gastroenteritis. *Epidemiology and infection*, 125(1):93–8.

Cheetham, S., Souza, M., Meulia, T., Grimes, S., Han, M. G., and Saif, L. J. (2006). Pathogenesis of a genogroup II human norovirus in gnotobiotic pigs. *Journal of virology*, 80(21):10372–81.

Choi, Y. S., Koo, E. S., Kim, M. S., Choi, J. D., Shin, Y., and Jeong, Y. S. (2017). Re-emergence of a GII.4 Norovirus Sydney 2012 Variant Equipped with GII.P16 RdRp and Its Predominance over Novel Variants of GII.17 in South Korea in 2016. *Food and Environmental Virology*, 9(2):168–178.

Chung, L., Bailey, D., Leen, E. N., Emmott, E. P., Chaudhry, Y., Roberts, L. O., Curry, S., Locker, N., and Goodfellow, I. G. (2014). Norovirus translation requires an interaction between the C terminus of the genome-linked viral protein VPg and eukaryotic translation initiation factor 4G. *Journal of Biological Chemistry*, 289(31):21738–21750.

Cortes-Penfield, N. W., Ramani, S., Estes, M. K., and Atmar, R. L. (2017). Prospects and Challenges in the Development of a Norovirus Vaccine.

Currier, R. L., Payne, D. C., Staat, M. A., Selvarangan, R., Shirley, S. H., Halasa, N., Boom, J. A., Englund, J. A., Szilagyi, P. G., Harrison, C. J., Klein, E. J., Weinberg, G. A., Wikswo, M. E., Parashar, U., Vinjé, J., and Morrow, A. L. (2015). Innate susceptibility to norovirus infections influenced by FUT2 genotype in a United States pediatric population. *Clinical Infectious Diseases*, 60(11):1631–1638.

da Silva Poló, T., Peiró, J. R., Mendes, L. C. N., Ludwig, L. F., de Oliveira-Filho, E. F., Bucardo, F., Huynen, P., Melin, P., Thiry, E., and Mauroy, A. (2016). Human norovirus infection in Latin America.

de Graaf, M., Bodewes, R., van Elk, C. E., van de Bildt, M., Getu, S., Aron, G. I., Verjans, G. M. G. M., Osterhaus, A. D. M. E., van den Brand, J. M. A., Kuiken, T., and Koopmans, M. P. G. (2017). Norovirus infection in harbor porpoises. *Emerging Infectious Diseases*, 23(1):87–91.

de Graaf, M., van Beek, J., and Koopmans, M. P. G. (2016). Human norovirus transmission and evolution in a changing world. *Nature Reviews Microbiology*, 14(7):421–433.

De Graaf, M., Van Beek, J., Vennema, H., Podkolzin, A. T., Hewitt, J., Bucardo, F., Templeton, K., Mans, J., Nordgren, J., Reuter, G., Lynch, M., Rasmussen, L. D., Iritani, N., Chan, M. C., Martella, V., Ambert-Balay, K., Vinj??, J., White, P. A., and Koopmans, M. P. (2015). Emergence of a novel GII.17 norovirus ??? end of the GII.4 era? *Eurosurveillance*, 20(26):1–8.

Debbink, K., Donaldson, E. F., Lindesmith, L. C., and Baric, R. S. (2012a). Genetic Mapping of a Highly Variable Norovirus GII.4 Blockade Epitope: Potential Role in Escape from Human Herd Immunity. *Journal of Virology*, 86(2):1214–1226.

Debbink, K., Lindesmith, L. C., Donaldson, E. F., Baric, R. S., and Rosenbaum, P. (2012b). Norovirus Immunity and the Great Escape. *PLoS Pathogens*, 8(10):e1002921.

Debbink, K., Lindesmith, L. C., Donaldson, E. F., Costantini, V., Beltramello, M., Corti, D., Swanstrom, J., Lanzavecchia, A., Vinjé, J., and Baric, R. S. (2013). Emergence of

new pandemic gII.4 sydney norovirus strain correlateswith escape from herd immunity. *Journal of Infectious Diseases*, 208(11):1877–1887.

Debbink, K., Lindesmith, L. C., Ferris, M. T., Swanstrom, J., Beltramello, M., Corti, D., Lanzavecchia, A., and Baric, R. S. (2014). Within-host evolution results in antigenically distinct GII.4 noroviruses. *Journal of virology*, 88(13):7244–55.

Donaldson, E. F., Lindesmith, L. C., Lobue, A. D., and Baric, R. S. (2008). Norovirus pathogenesis: Mechanisms of persistence and immune evasion in human populations.

Donaldson, E. F., Lindesmith, L. C., Lobue, A. D., and Baric, R. S. (2010). Viral shapeshifting: norovirus evasion of the human immune system. *Nature reviews. Microbiology*, 8(3):231–241.

Eden, J.-S., Chisholm, R. H., Bull, R. A., White, P. A., Holmes, E. C., and Tanaka, M. M. (2017). Persistent infections in immunocompromised hosts are rarely sources of new pathogen variants. *Virus Evolution*, 3(2):219–22.

Eden, J. S., Hewitt, J., Lim, K. L., Boni, M. F., Merif, J., Greening, G., Ratcliff, R. M., Holmes, E. C., Tanaka, M. M., Rawlinson, W. D., and White, P. A. (2014). The emergence and evolution of the novel epidemic norovirus GII.4 variant Sydney 2012. *Virology*, 450-451:106–113.

Eden, J.-S., Tanaka, M. M., Boni, M. F., Rawlinson, W. D., and White, P. A. (2013). Recombination within the pandemic norovirus GII.4 lineage. *Journal of virology*, 87(11):6270–82.

Edgar, R. C. (2008). Muscle. *BMC Bioinformatics*.

Emmott, E., de Rougemont, A., Haas, J., and Goodfellow, I. (2017a). Spatial and temporal control of norovirus protease activity is determined by polyprotein processing and intermolecular interactions within the viral replication complex. *bioRxiv*.

Emmott, E., Sorgeloos, F., Caddy, S. L., and Heesom, K. (2017b). Norovirus-mediated modification of the translational landscape via virus and host-induced cleavage of translation initiation factors. *Molecular and Cellular Proteomics*, pages 1–32.

Ettayebi, K., Crawford, S. E., Murakami, K., Broughman, J. R., Karandikar, U., Tenge, V. R., Neill, F. H., Blutt, S. E., Zeng, X.-l., Qu, L., Kou, B., Antone, R., Burrin, D., Graham, D. Y., Ramani, S., Atmar, R. L., and Mary, K. (2016). Replication of human noroviruses in stem cell derived human enteroids. *Science*, 5211(August):1–12.

Firth, A. E. and Brierley, I. (2012). Non-canonical translation in RNA viruses.

Fu, J., Ai, J., Jin, M., Jiang, C., Zhang, J., Shi, C., Lin, Q., Yuan, Z., Qi, X., Bao, C., Tang, F., and Zhu, Y. (2015). Emergence of a new GII.17 norovirus variant in patients with acute gastroenteritis in jiangsu, China, september 2014 to march 2015. *Eurosurveillance*, 20(24):1–7.

Furman, L. M., Maaty, W. S., Petersen, L. K., Ettayebi, K., Hardy, M. E., and Bothner, B. (2009). Cysteine protease activation and apoptosis in Murine norovirus infection. *Virology journal*, 6:139.

Galeano, M. E., Martinez, M., Amarilla, A. A., Russomando, G., Miagostovich, M. P., Parra, G. I., and Leite, J. P. (2013). Molecular epidemiology of norovirus strains in Paraguayan children during 2004-2005: Description of a possible new GII.4 cluster. *Journal of Clinical Virology*, 58(2):378–384.

Gallimore, C. I., Taylor, C., Gennery, A. R., Cant, A. J., Galloway, A., Iturriza-Gomara, M., and Gray, J. J. (2006). Environmental monitoring for gastroenteric viruses in a pediatric primary immunodeficiency unit. *Journal of Clinical Microbiology*, 44(2):395–399.

Gallimore, C. I., Taylor, C., Gennery, A. R., Cant, A. J., Galloway, A., Xerry, J., Adigwe, J., and Gray, J. J. (2008). Contamination of the hospital environment with gastroenteric viruses: Comparison of two pediatric wards over a winter season. *Journal of Clinical Microbiology*, 46(9):3112–3115.

GBD 2013 Mortality and Causes of Death Collaborators (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet (London, England)*, 385(9963):117–71.

Giammanco, G. M., De Grazia, S., Tummolo, F., Bonura, F., Calderaro, A., Buonavoglia, A., Martella, V., and Medici, M. C. (2013). Norovirus GII.4/Sydney/2012 in Italy, Winter 20122013. *Emerging Infectious Diseases*, 19(8):1348–1349.

Glass, P. J., Zeng, C. Q., and Estes, M. K. (2003). Two Nonoverlapping Domains on the Norwalk Virus Open Reading Frame 3 (ORF3) Protein Are Involved in the Formation of the Phosphorylated 35K Protein and in ORF3-Capsid Protein Interactions. *Journal of Virology*, 77(6):3569–3577.

Green, K. (2007). Caliciviridae: The noroviruses. *Fields Virology*, 2(v. 1):3177.

Green, K. Y. (2016). Editorial Commentary : Noroviruses and B Cells. *Clinical Infectious Diseases*, 62(9):1139–1140.

Green, K. Y., Mory, a., Fogg, M. H., Weisberg, a., Belliot, G., Wagner, M., Mitra, T., Ehrenfeld, E., Cameron, C. E., and Sosnovtsev, S. V. (2002). Isolation of enzymatically active replication complexes from feline calicivirus-infected cells. *J.Virol.*, 76(17):8582–8595.

Haga, K., Fujimoto, A., Takai-Todaka, R., Miki, M., Doan, Y. H., Murakami, K., Yokoyama, M., Murata, K., Nakanishi, A., and Katayama, K. (2016). Functional receptor molecules CD300lf and CD300ld within the CD300 family enable murine noroviruses to infect cells. *Proceedings of the National Academy of Sciences of the United States of America*, page 201605575.

Hall, A. J. (2012). Noroviruses: The Perfect Human Pathogens? *Journal of Infectious Diseases*, 205(11):1622–1624.

Hansman, G. S., Biertumpfel, C., Georgiev, I., McLellan, J. S., Chen, L., Zhou, T., Katayama, K., and Kwong, P. D. (2011). Crystal Structures of GII.10 and GII.12 Norovirus Protruding Domains in Complex with Histo-Blood Group Antigens Reveal Details for a Potential Site of Vulnerability. *Journal of Virology*, 85(13):6687–6701.

Harris, J. P., Iturriza-Gomara, M., and O'Brien, S. J. (2017). Re-assessing the total burden of norovirus circulating in the United Kingdom population. *Vaccine*, 35(6):853–855.

Hasing, M. E., Lee, B. E., Preiksaitis, J. K., Tellier, R., Honish, L., Senthilselvan, A., and Pang, X. L. (2013). Emergence of a new norovirus GII.4 variant and changes in the historical biennial pattern of norovirus outbreak activity in Alberta, Canada, from 2008 to 2013. *Journal of Clinical Microbiology*, 51(7):2204–2211.

Hickman, D., Jones, M. K., Zhu, S., Kirkpatrick, E., Ostrov, D. A., Wang, X., Ukhanova, M., Sun, Y., Mai, V., Salemi, M., and Karst, S. M. (2014). The effect of malnutrition on norovirus infection. *mBio*, 5(2).

Högbom, M., Jäger, K., Robel, I., Unge, T., and Rohayem, J. (2009). The active form of the norovirus RNA-dependent RNA polymerase is a homodimer with cooperative activity. *Journal of General Virology*, 90(2):281–291.

Hyde, J. L. and Mackenzie, J. M. (2010). Subcellular localization of the MNV-1 ORF1 proteins and their potential roles in the formation of the MNV-1 replication complex. *Virology*, 406(1):138–148.

Hyde, J. L., Sosnovtsev, S. V., Green, K. Y., Wobus, C., Virgin, H. W., and Mackenzie, J. M. (2009). Mouse Norovirus Replication Is Associated with Virus-Induced Vesicle Clusters Originating from Membranes Derived from the Secretory Pathway. *Journal of Virology*, 83(19):9709–9719.

Inns, T., Harris, J., Vivancos, R., Iturriza-Gomara, M., and O'Brien, S. (2017). Community-based surveillance of norovirus disease: a systematic review. *BMC Infectious Diseases*, 17(1):657.

Jiang, X., Wang, M., Graham, D. Y., and Estes, M. K. (1992). Expression, self-assembly, and antigenicity of the Norwalk virus capsid protein. *Journal of virology*, 66(11):6527–32.

Jones, M. K., Grau, K. R., Costantini, V., Kolawole, A. O., de Graaf, M., Freiden, P., Graves, C. L., Koopmans, M., Wallet, S. M., Tibbetts, S. A., Schultz-Cherry, S., Wobus, C. E., Vinjé, J., and Karst, S. M. (2015). Human norovirus culture in B cells. *Nature Protocols*, 10(12):1939–1947.

Jones, M. K., Watanabe, M., Zhu, S., Graves, C. L., Keyes, L. R., Grau, K. R., Gonzalez-Hernandez, M. B., Iovine, N. M., Wobus, C. E., Vinje, J., Tibbetts, S. A., Wallet,

S. M., and Karst, S. M. (2014). Enteric bacteria promote human and mouse norovirus infection of B cells. *Science*, 346(6210):755–759.

Kaiser, W. J., Chaudhry, Y., Sosnovtsev, S. V., and Goodfellow, I. G. (2006). Analysis of protein-protein interactions in the feline calicivirus replication complex. *The Journal of general virology*, 87(Pt 2):363–8.

Kambhampati, A., Payne, D. C., Costantini, V., and Lopman, B. A. (2015). Host Genetic Susceptibility to Enteric Viruses: A Systematic Review and Metaanalysis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 62:civ873–.

Kamel, A. H., Ali, M. A., El-Nady, H. G., De Rougemont, A., Pothier, P., and Belliot, G. (2009). Predominance and circulation of enteric viruses in the region of greater Cairo, Egypt. *Journal of Clinical Microbiology*, 47(4):1037–1045.

Kapikian, A. Z., Wyatt, R. G., Dolin, R., Thornhill, T. S., Kalica, A. R., and Chanock, R. M. (1972). Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. *Journal of virology*, 10(5):1075–1081.

Karandikar, U. C., Crawford, S. E., Ajami, N. J., Murakami, K., Kou, B., Ettayebi, K., Papanicolaou, G. A., Jongwutiwes, U., Perales, M. A., Shia, J., Mercer, D., Finegold, M. J., Vinjé, J., Atmar, R. L., and Estes, M. K. (2016). Detection of human norovirus in intestinal biopsies from immunocompromised transplant patients. *Journal of General Virology*, 97(9):2291–2300.

Karst, S. M. (2003). STAT1-Dependent Innate Immunity to a Norwalk-Like Virus. *Science*, 299(5612):1575–1578.

Karst, S. M. and Baric, R. S. (2015). What is the reservoir of emergent human norovirus strains? *Journal of virology*, 89(11):5756–9.

Karst, S. M., Wobus, C. E., Goodfellow, I. G., Green, K. Y., and Virgin, H. W. (2014). Advances in norovirus biology.

Kim, H. S., Hyun, J., Kim, H. S., Kim, J. S., Song, W., and Lee, K. M. (2013). Emergence of GII.4 Sydney norovirus in South Korea during the winter of 2012-2013. *Journal of Microbiology and Biotechnology*, 23(11):1641–1643.

Kindberg, E. and Svensson, L. (2009). Genetic basis of host resistance to norovirus infection. *Future Virology*, 4(4):369–382.

Kirk, M. D., Pires, S. M., Black, R. E., Caipo, M., Crump, J. A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C. L., Hald, T., Hall, A. J., Keddy, K. H., Lake, R. J., Lanata, C. F., Torgerson, P. R., Havelaar, A. H., and Angulo, F. J. (2015). World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS Medicine*, 12(12).

Kiulia, N. M., Mans, J., Mwenda, J. M., and Taylor, M. B. (2014). Norovirus GII.17 Predominates in Selected Surface Water Sources in Kenya. *Food and Environmental Virology*, 6(4):221–231.

Koehl, P. and Levitt, M. (1999). Structure-based conformational preferences of amino acids. *Proceedings of the National Academy of Sciences*, 96(22):12524–12529.

Koromyslova, A. D., Leuthold, M. M., Bowler, M. W., and Hansman, G. S. (2015). The sweet quartet: Binding of fucose to the norovirus capsid. *Virology*, 483:203–208.

Kosakovsky Pond, S. L., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679.

Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., Wu, Y., Sow, S. O., Sur, D., Breiman, R. F., Faruque, A. S., Zaidi, A. K., Saha, D., Alonso, P. L., Tamboura, B., Sanogo, D., Onwuchekwa, U., Manna, B., Ramamurthy, T., Kanungo, S., Ochieng, J. B., Omore, R., Oundo, J. O., Hossain, A., Das, S. K., Ahmed, S., Qureshi, S., Quadri, F., Adegbola, R. A., Antonio, M., Hossain, M. J., Akinsola, A., Mandomando, I., Nhampossa, T., Acácio, S., Biswas, K., O'Reilly, C. E., Mintz, E. D., Berkeley, L. Y., Muhsen, K., Sommerfelt, H., Robins-Browne, R. M., and Levine, M. M. (2013). Burden and aetiology of diarrhoeal disease in infants and

young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. *The Lancet*, 382(9888):209–222.

Kroneman, A., Vega, E., Vennema, H., Vinjé, J., White, P. A., Hansman, G., Green, K., Martella, V., Katayama, K., and Koopmans, M. (2013). Proposal for a unified norovirus nomenclature and genotyping. *Archives of Virology*, 158(10):2059–2068.

Kroneman, A., Vennema, H., Deforche, K., Avoort, H., Peñaranda, S., Oberste, M. S., Vinjé, J., and Koopmans, M. (2011). An automated genotyping tool for enteroviruses and noroviruses. *Journal of Clinical Virology*, 51(2):121–125.

Kroneman, A., Verhoef, L., Harris, J., Vennema, H., Duizer, E., Van Duynhoven, Y., Gray, J., Iturriza, M., Böttiger, B., Falkenhorst, G., Johnsen, C., Von Bonsdorff, C. H., Maunula, L., Kuusi, M., Pothier, P., Gallay, A., Schreier, E., Höhne, M., Koch, J., Szücs, G., Reuter, G., Krisztalovics, K., Lynch, M., McKeown, P., Foley, B., Coughlan, S., Ruggeri, F. M., Di Bartolo, I., Vainio, K., Isakbaeva, E., Poljsak-Prijatelj, M., Hocevar Grom, A., Zimsek Mijovski, J., Bosch, A., Buesa, J., Sanchez Fauquier, A., Hernandéz-Pezzi, G., Hedlund, K. O., and Koopmans, M. (2008). Analysis of integrated virological and epidemiological reports of norovirus outbreaks collected within the Foodborne Viruses in Europe network from 1 July 2001 to 30 June 2006. *Journal of Clinical Microbiology*, 46(9):2959–2965.

Kumazaki, M. and Usuku, S. (2015). Genetic analysis of norovirus GII.4 variant strains detected in outbreaks of gastroenteritis in Yokohama, Japan, from the 2006-2007 to the 2013-2014 seasons.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132.

Lay, M. K., Atmar, R. L., Guix, S., Bharadwaj, U., He, H., Neill, F. H., Sastry, K. J., Yao, Q., and Estes, M. K. (2010). Norwalk virus does not replicate in human macrophages or dendritic cells derived from the peripheral blood of susceptible humans. *Virology*, 406(1):1–11.

Le Pendu, J., Ruvoën-Clouet, N., Kindberg, E., and Svensson, L. (2006). Mendelian resistance to human norovirus infections.

Lee, C. C., Feng, Y., Chen, S. Y., Tsai, C. N., Lai, M. W., and Chiu, C. H. (2015). Emerging Norovirus GII.17 in Taiwan.

Lee, R. M., Lessler, J., Lee, R. A., Rudolph, K. E., Reich, N. G., Perl, T. M., and Cummings, D. A. (2013). Incubation periods of viral gastroenteritis: a systematic review. *BMC Infectious Diseases*, 13(1):446.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., and Suchard, M. A. (2014). Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens*, 10(2).

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9).

Li, L., Shan, T., Wang, C., Cote, C., Kolman, J., Onions, D., Gulland, F. M. D., and Delwart, E. (2011). The Fecal Viral Flora of California Sea Lions. *Journal of Virology*, 85(19):9909–9917.

Lindesmith, L., Moe, C., Le pendu, J., Frelinger, J. A., Treanor, J., and Baric, R. S. (2005). Cellular and humoral immunity following Snow Mountain virus challenge. *Journal of virology*, 79(5):2900–9.

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. (2003). Human susceptibility and resistance to Norwalk virus infection. *Nature Medicine*, 9(5):548–553.

Lindesmith, L. C., Beltramello, M., Donaldson, E. F., Corti, D., Swanstrom, J., Debbink, K., Lanzavecchia, A., and Baric, R. S. (2012a). Immunogenetic mechanisms driving norovirus GII.4 antigenic variation. *PLoS Pathogens*, 8(5).

Lindesmith, L. C., Costantini, V., Swanstrom, J., Debbink, K., Donaldson, E. F., Vinjé, J., and Baric, R. S. (2013). Emergence of a norovirus GII.4 strain correlates with changes in evolving blockade epitopes. *Journal of virology*, 87(5):2803–13.

Lindesmith, L. C., Debbink, K., Swanstrom, J., Vinjé, J., Costantini, V., Baric, R. S., and

Donaldson, E. F. (2012b). Monoclonal antibody-based antigenic mapping of norovirus GII.4-2002. *Journal of virology*, 86(2):873–83.

Lindesmith, L. C., Donaldson, E., Leon, J., Moe, C. L., Frelinger, J. a., Johnston, R. E., Weber, D. J., and Baric, R. S. (2010). Heterotypic humoral and cellular immune responses following Norwalk virus infection. *Journal of virology*, 84(4):1800–1815.

Lindesmith, L. C., Donaldson, E. F., and Baric, R. S. (2011). Norovirus GII.4 Strain Antigenic Variation. *Journal of Virology*, 85(1):231–242.

Lindesmith, L. C., Donaldson, E. F., Beltramello, M., Pintus, S., Corti, D., Swanstrom, J., Debbink, K., Jones, T. A., Lanzavecchia, A., and Baric, R. S. (2014). Particle Conformation Regulates Antibody Access to a Conserved GII.4 Norovirus Blockade Epitope. *Journal of Virology*, 88(16):8826–8842.

Lindesmith, L. C., Donaldson, E. F., LoBue, A. D., Cannon, J. L., Zheng, D. P., Vinje, J., and Baric, R. S. (2008). Mechanisms of GII.4 norovirus persistence in human populations. *PLoS Medicine*, 5(2):0269–0290.

Lindesmith, L. C., Kocher, J. F., Donaldson, E. F., Debbink, K., Mallory, M. L., Swann, E. W., Brewer-Jenson, P. D., and Baric, R. S. (2017a). Emergence of Novel Human Norovirus GII.17 Strains Correlates with Changes in Blockade Antibody Epitopes. *The Journal of Infectious Diseases*.

Lindesmith, L. C., Mallory, M. L., Jones, T. A., Richardson, C., Goodwin, R. R., Baehner, F., Mendelman, P. M., Bargatze, R. F., and Baric, R. S. (2017b). Impact of pre-exposure history and host genetics on antibody avidity following norovirus vaccination. *Journal of Infectious Diseases*, 215(6):984–991.

Liu, B. L., Lambden, P. R., Günther, H., Otto, P., Elschner, M., and Clarke, I. N. (1999). Molecular characterization of a bovine enteric calicivirus: relationship to the Norwalk-like viruses. *Journal of virology*, 73(1):819–25.

Liu, L. T.-C., Kuo, T.-Y., Wu, C.-Y., Liao, W.-T., Hall, A. J., and Wu, F.-T. (2017). Recombinant GII.P16-GII.2 Norovirus, Taiwan, 2016. *Emerging Infectious Diseases*, 23(7):1180–1183.

Liu, P., Wang, X., Lee, J.-C., Teunis, P., Hu, S., Paradise, H. T., and Moe, C. (2014). Genetic Susceptibility to Norovirus GII.3 and GII.4 Infections in Chinese Pediatric Diarrheal Disease. *The Pediatric Infectious Disease Journal*, 33(11):e305–e309.

Lopman, B., Gastañaduy, P., Park, G. W., Hall, A. J., Parashar, U. D., and Vinjé, J. (2012). Environmental transmission of norovirus gastroenteritis. *Current Opinion in Virology*, 2(1):96–102.

Lopman, B., Vennema, H., Kohli, E., Pothier, P., Sanchez, A., Negredo, A., Buesa, J., Schreier, E., Reacher, M., Brown, D., Gray, J., Iturriza, M., Gallimore, C., Bottiger, B., Hedlund, K. O., Torv??n, M., Von Bonsdorff, C. H., Maunula, L., Poljsak-Prijatelj, M., Zimsek, J., Reuter, G., Sz??cs, G., Melegh, B., Svennson, L., Van Duijnhoven, Y., and Koopmans, M. (2004a). Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant.

Lopman, B. a., Reacher, M. H., Vipond, I. B., Sarangi, J., and Brown, D. W. G. (2004b). Clinical manifestation of norovirus gastroenteritis in health care settings. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 39(3):318–324.

Lopman, B. A., Steele, D., Kirkwood, C. D., and Parashar, U. D. (2016). The Vast and Varied Global Burden of Norovirus: Prospects for Prevention and Control.

Lopman, B. A., Trivedi, T., Vicuña, Y., Costantini, V., Collins, N., Gregoricus, N., Parashar, U., Sandoval, C., Broncano, N., Vaca, M., Chico, M. E., Vinjé, J., and Cooper, P. J. (2015). Norovirus infection and disease in an ecuadorian birth cohort: Association of certain norovirus genotypes with host FUT2 secretor status. *Journal of Infectious Diseases*, 211(11):1813–1821.

Lu, J., Fang, L., Sun, L., Zeng, H., Li, Y., Zheng, H., Wu, S., Yang, F., Song, T., Lin, J., Ke, C., Zhang, Y., Vinjé, J., and Li, H. (2017). Association of GII.P16-GII.2 Recombinant Norovirus Strain with Increased Norovirus Outbreaks, Guangdong, China, 2016. *Emerging infectious diseases*, 23(7):1188–1190.

Lu, J., Fang, L., Zheng, H., Lao, J., Yang, F., Sun, L., Xiao, J., Lin, J., Song, T., Ni, T., Raghwani, J., Ke, C., Faria, N. R., Bowden, T. A., Pybus, O. G., and Li, H. (2016).

The evolution and transmission of epidemic gii.17 noroviruses. *Journal of Infectious Diseases*, 214(4):556–564.

Lu, J., Sun, L., Fang, L., Yang, F., Mo, Y., Lao, J., Zheng, H., Tan, X., Lin, H., Rutherford, S., Guo, L., Ke, C., and Hui, L. (2015). Gastroenteritis outbreaks caused by norovirus GII.17, Guangdong Province, China, 20142015. *Emerging Infectious Diseases*, 21(7):1240–1242.

Malm, M., Tamminen, K., Lappalainen, S., Uusi-Kerttula, H., Vesikari, T., and Blazevic, V. (2015). Genotype considerations for virus-like particle-based bivalent norovirus vaccine composition. *Clinical and Vaccine Immunology*, 22(6):656–663.

Malm, M., Uusi-Kerttula, H., Vesikari, T., and Blazevic, V. (2014). High serum levels of norovirus genotype-specific blocking antibodies correlate with protection from infection in children. *Journal of Infectious Diseases*, 210(11):1755–1762.

Mans, J., Armah, G. E., Steele, A. D., and Taylor, M. B. (2016). Norovirus epidemiology in Africa: A review.

Mans, J., de Villiers, J. C., du Plessis, N. M., Avenant, T., and Taylor, M. B. (2010). Emerging norovirus GII.4 2008 variant detected in hospitalised paediatric patients in South Africa. *Journal of Clinical Virology*, 49(4):258–264.

Mans, J., Murray, T. Y., Nadan, S., Netshikweta, R., Page, N. A., and Taylor, M. B. (2015). Norovirus diversity in children with gastroenteritis in South Africa from 2009 to 2013: GII.4 variants and recombinant strains predominate. *Epidemiology and Infection*, 22:1–10.

Martella, V., Decaro, N., Lorusso, E., Radogna, A., Moschidou, P., Amorisco, F., Lucente, M. S., Desario, C., Mari, V., Elia, G., Banyai, K., Carmichael, L. E., and Buonavoglia, C. (2009). Genetic heterogeneity and recombination in canine noroviruses. *Journal of virology*, 83(21):11391–11396.

Martella, V., Lorusso, E., Decaro, N., Elia, G., Radogna, A., D'Abramo, M., Desario, C., Cavalli, A., Corrente, M., Camero, M., Germinario, C. A., Bányai, K., Di Martino, B., Marsilio, F., Carmichael, L. E., and Buonavoglia, C. (2008). Detection and molecular

characterization of a canine norovirus. *Emerging Infectious Diseases*, 14(8):1306–1308.

Matsushima, Y., Ishikawa, M., Shimizu, T., Komane, A., Kasuo, S., Shinohara, M., Nagasawa, K., Kimura, H., Ryo, A., Okabe, N., Haga, K., Doan, Y., Katayama, K., and Shimizu, H. (2015). Genetic analyses of GII.17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. *Eurosurveillance*, 20(26):21173.

Matsushima, Y., Shimizu, T., Ishikawa, M., Komane, A., Okabe, N., Ryo, A., Kimura, H., Katayama, K., and Shimizu, H. (2016). Complete Genome Sequence of a Recombinant GII.P16-GII.4 Norovirus Detected in Kawasaki City, Japan, in 2016. *Genome Announcements*, 4(5):e01099–16.

Mattison, K., Shukla, A., Cook, A., Pollari, F., Friendship, R., Kelton, D., Bidawid, S., and Farber, J. M. (2007). Human noroviruses in swine and cattle. *Emerging Infectious Diseases*, 13(8):1184–1188.

McCartney, S. A., Thackray, L. B., Gitlin, L., Gilfillan, S., Virgin IV, H. W., and Colonna, M. (2008). MDA-5 recognition of a murine norovirus. *PLoS Pathogens*, 4(7).

McCune, B. T., Tang, W., Lu, J., Eaglesham, J. B., Thorne, L., Mayer, A. E., Condiff, E., Nice, T. J., Goodfellow, I., Krezel, A. M., and Virgin, H. W. (2017). Noroviruses Co-opt the Function of Host Proteins VAPA and VAPB for Replication via a Phenylalanine-Phenylalanine-Acidic-Tract-Motif Mimic in Nonstructural Viral Protein NS1/2. *mBio*, 8(4):1–17.

McFadden, N., Bailey, D., Carrara, G., Benson, A., Chaudhry, Y., Shortland, A., Heeney, J., Yarovinsky, F., Simmonds, P., Macdonald, A., and Goodfellow, I. (2011). Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. *PLoS Pathogens*, 7(12).

Mercer, J., Schelhaas, M., and Helenius, A. (2010). Virus Entry by Endocytosis. *Annual Review of Biochemistry*, 79(1):803–833.

Mesquita, J. R., Barclay, L., Nascimento, M. S. J., and Vinjé, J. (2010). Novel norovirus in dogs with diarrhea. *Emerging Infectious Diseases*, 16(6):980–982.

Mori, K., Chu, P. Y., Motomura, K., Somura, Y., Nagano, M., Kimoto, K., Akiba, T., Kai, A., and Sadamasu, K. (2017a). Genomic analysis of the evolutionary lineage of Norovirus GII.4 from archival specimens during 1975???1987 in Tokyo. *Journal of Medical Virology*, 89(2):363–367.

Mori, K., Nagano, M., Kimoto, K., Somura, Y., Akiba, T., Hayashi, Y., Sadamasu, K., and Ka, A. (2017b). Detection of enteric viruses in fecal specimens from nonbacterial foodborne gastroenteritis outbreaks in Tokyo, Japan between 1966 and 1983. *Japanese Journal of Infectious Diseases*, 70(2):143–151.

Motomura, K., Yokoyama, M., Ode, H., Nakamura, H., Mori, H., Kanda, T., Oka, T., Katayama, K., Noda, M., Tanaka, T., Takeda, N., and Sato, H. (2010). Divergent evolution of norovirus GII/4 by genome recombination from May 2006 to February 2009 in Japan. *Journal of virology*, 84(16):8085–8097.

Napthine, S., Lever, R. A., Powell, M. L., Jackson, R. J., Brown, T. D. K., and Brierley, I. (2009). Expression of the VP2 protein of murine norovirus by a translation termination-reinitiation strategy. *PLoS ONE*, 4(12).

Ng, K. K. S., Pendás-Franco, N., Rojo, J., Boga, J. A., Machín, Á., Martín Alonso, J. M., and Parra, F. (2004). Crystal Structure of Norwalk Virus Polymerase Reveals the Carboxyl Terminus in the Active Site Cleft. *Journal of Biological Chemistry*, 279(16):16638–16645.

Nice, T. J., Strong, D. W., McCune, B. T., Pohl, C. S., Virgin, H. W., Thackray, L. B., Smith, T. J., and Virgin, H. W. (2013). A Single-Amino-Acid Change in Murine Norovirus NS1/2 Is Sufficient for Colonic Tropism and Persistence. *Journal of virology*, 86(6):327–334.

Niendorf, S., Jacobsen, S., Faber, M., Eis-Hübinger, A. M., Hofmann, J., Zimmermann, O., Höhne, M., and Bock, C. T. (2017). Steep rise in norovirus cases and emergence of a new recombinant strain GII.P16-GII.2, Germany, winter 2016. *Euro surveillance :*

bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 22(4).

Niendorf, S., Klemm, U., Marques, A. M., Bock, C. T., and Höhne, M. (2016). Infection with the persistent murine norovirus strain MNV-S99 suppresses IFN-Beta release and activation of stat1 In vitro. *PLoS ONE*, 11(6).

Noel, J. S., Fankhauser, R. L., Ando, T., Monroe, S. S., and Glass, R. I. (1999). Identification of a Distinct Common Strain of Norwalklike Viruses Having a Global Distribution. *The Journal of Infectious Diseases*, 179(6):1334–1344.

Nordgren, J., Kindberg, E., Lindgren, P. E., Matussek, A., and Svensson, L. (2010). Norovirus gastroenteritis outbreak with a secretor-independent susceptibility pattern, Sweden. *Emerging Infectious Diseases*, 16(1):81–87.

Nordgren, J., Nitiema, L. W., Ouermi, D., Simpore, J., and Svensson, L. (2013). Host Genetic Factors Affect Susceptibility to Norovirus Infections in Burkina Faso. *PLoS ONE*, 8(7).

Nordgren, J., Sharma, S., Kambhampati, A., Lopman, B., Svensson, L., and Lynfield, R. (2016). Innate Resistance and Susceptibility to Norovirus Infection. *PLOS Pathogens*, 12(4):e1005385.

O'Brien, S. J., Donaldson, A. L., Iturriza-Gomara, M., and Tam, C. C. (2016). Age-Specific Incidence Rates for Norovirus in the Community and Presenting to Primary Healthcare Facilities in the United Kingdom. *Journal of Infectious Diseases*, 213:S15–S18.

Oliver, S. L., Dastjerdi, A. M., Wong, S., El-Attar, L., Gallimore, C., Brown, D. W. G., Green, J., and Bridger, J. C. (2003). Molecular Characterization of Bovine Enteric Caliciviruses: a Distinct Third Genogroup of Noroviruses (Norwalk-Like Viruses) Unlikely To Be of Risk to Humans. *Journal of Virology*, 77(4):2789–2798.

Orchard, R. C., Wilen, C. B., Doench, J. G., Baldridge, M. T., McCune, B. T., Lee, Y.-C. J., Lee, S., Pruett-Miller, S. M., Nelson, C. A., Fremont, D. H., and Virgin, H. W. (2016). Discovery of a proteinaceous cellular receptor for a norovirus. *Science*, 353(6302):933–936.

Otto, P. H., Clarke, I. N., Lambden, P. R., Salim, O., Reetz, J., and Liebler-Tenorio, E. M. (2011). Infection of Calves with Bovine Norovirus GIII.1 Strain Jena Virus: an Experimental Model To Study the Pathogenesis of Norovirus Infection. *Journal of Virology*, 85(22):12013–12021.

Park, S. I., Jeong, C., Kim, H. H., Park, S. H., Park, S. J., Hyun, B. H., Yang, D. K., Kim, S. K., Kang, M. I., and Cho, K. O. (2007). Molecular epidemiology of bovine noroviruses in South Korea. *Veterinary Microbiology*, 124(1-2):125–133.

Parra, G. I., Squires, R. B., Karangwa, C. K., Johnson, J. A., Lepore, C. J., Sosnovtsev, S. V., and Green, K. Y. (2017). Static and Evolving Norovirus Genotypes: Implications for Epidemiology and Immunity. *PLoS Pathogens*, 13(1).

Parrino, T. A., Schreiber, D. S., Trier, J. S., Kapikian, A. Z., and Blacklow, N. R. (1977). Clinical immunity in acute gastroenteritis caused by Norwalk agent. *The New England journal of medicine*, 297(2):86–9.

Patel, M. M., Hall, A. J., Vinjé, J., and Parashar, U. D. (2009). Noroviruses: A comprehensive review.

Patel, M. M., Widdowson, M. A., Glass, R. I., Akazawa, K., Vinjé, J., and Parashar, U. D. (2008). Systematic literature review of role of noroviruses in sporadic gastroenteritis.

Payne, D. C., Vinjé, J., Szilagyi, P. G., Edwards, K. M., Staat, M. A., Weinberg, G. A., Hall, C. B., Chappell, J., Bernstein, D. I., Curns, A. T., Wikswo, M., Shirley, S. H., Hall, A. J., Lopman, B., and Parashar, U. D. (2013). Norovirus and Medically Attended Gastroenteritis in U.S. Children. *New England Journal of Medicine*, 368(12):1121–1130.

Pires, S. M., Fischer-Walker, C. L., Lanata, C. F., Devleesschauwer, B., Hall, A. J., Kirk, M. D., Duarte, A. S., Black, R. E., and Angulo, F. J. (2015). Aetiology-specific estimates of the global and regional incidence and mortality of diarrhoeal diseases commonly transmitted through food. *PLoS ONE*, 10(12).

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*, 23(10):1891–1901.

Prasad, B. V., Hardy, M. E., Dokland, T., Bella, J., Rossmann, M. G., and Estes, M. K. (1999). X-ray crystallographic structure of the Norwalk virus capsid. *Science*, 286(October):287–290.

Qu, L., Murakami, K., Broughman, J. R., Lay, M. K., Guix, S., Tenge, V. R., Atmar, R. L., and Estes, M. K. (2016). Replication of Human Norovirus RNA in Mammalian Cells Reveals Lack of Interferon Response. *Journal of Virology*, 90(19):8906–8923.

Ramani, S., Estes, M. K., and Atmar, R. L. (2016). Correlates of Protection against Norovirus Infection and DiseaseWhere Are We Now, Where Do We Go?

Ramani, S., Neill, F. H., Opekun, A. R., Gilger, M. A., Graham, D. Y., Estes, M. K., and Atmar, R. L. (2015). Mucosal and Cellular Immune Responses to Norwalk Virus. In *Journal of Infectious Diseases*, volume 212, pages 397–405.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1):vew007.

Reeck, A., Kavanagh, O., Estes, M. K., Opekun, A. R., Gilger, M. a., Graham, D. Y., and Atmar, R. L. (2010). Serological correlate of protection against norovirus-induced gastroenteritis. *The Journal of infectious diseases*, 202(8):1212–1218.

Robilotti, E., Deresinski, S., and Pinsky, B. A. (2015). Norovirus. *Clinical Microbiology Reviews*, 28(1):134–164.

Rockx, B., Baric, R. S., De Grijs, I., Duizer, E., and Koopmans, M. P. G. (2005). Characterization of the homo- and heterotypic immune responses after natural norovirus infection. *Journal of Medical Virology*, 77(3):439–446.

Rohayem, J., Robel, I., Jager, K., Scheffler, U., and Rudolph, W. (2006). Protein-Primed and De Novo Initiation of RNA Synthesis by Norovirus 3Dpol. *Journal of Virology*, 80(14):7060–7069.

Roth, A. N. and Karst, S. M. (2016). Norovirus mechanisms of immune antagonism.

Ruvoën-Clouet, N., Belliot, G., and Le Pendu, J. (2013). Noroviruses and histo-blood groups: The impact of common host genetic polymorphisms on virus transmission and evolution.

Sato, T., Stange, D. E., Ferrante, M., Vries, R. G. J., Van Es, J. H., Van Den Brink, S., Van Houdt, W. J., Pronk, A., Van Gorp, J., Siersema, P. D., and Clevers, H. (2011). Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology*, 141(5):1762–1772.

Schuit, M., Miller, D. M., Reddick-Elick, M. S., Wlazlowski, C. B., Filone, C. M., Herzog, A., Colf, L. A., Wahl-Jensen, V., Hevey, M., and Noah, J. W. (2016). Differences in the comparative stability of ebola virus makona-c05 and yambuku-mayinga in blood. *PLoS ONE*, 11(2).

Sdiri-Loulizi, K., Ambert-Balay, K., Gharbi-Khelifi, H., Sakly, N., Hassine, M., Chouchane, S., Guediche, M. N., Pothier, P., and Aouni, M. (2009). Molecular epidemiology of norovirus gastroenteritis investigated using samples collected from children in Tunisia during a four-year period: Detection of the norovirus variant GGII.4 hunter as early as January 2003. *Journal of Clinical Microbiology*, 47(2):421–429.

Seitz, S. R., Leon, J. S., Schwab, K. J., Lyon, G. M., Dowd, M., McDaniels, M., Abdulhafid, G., Fernandez, M. L., Lindesmith, L. C., Baric, R. S., and Moe, C. L. (2011). Norovirus infectivity in humans and persistence in water. *Applied and Environmental Microbiology*, 77(19):6884–6888.

Sharma, S., Carlsson, B., Czakó, R., Vene, S., Haglund, M., Ludvigsson, J., Larson, G., Hammarström, L., Sosnovtsev, S., Atmar, R., Green, K., Estes, M., and Svensson, L. (2017). Human sera collected between 1979 and 2010 possess blocking-antibody titers to pandemic GII.4 noroviruses isolated over three decades. *Journal of Virology*, 91(14).

Siebenga, J. J., Lemey, P., Pond, S. L. K., Rambaut, A., Vennema, H., and Koopmans, M. (2010). Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. *PLoS Pathogens*, 6(5):1–13.

Siebenga, J. J., Vennema, H., Renckens, B., de Bruin, E., van der Veer, B., Siezen, R. J.,

and Koopmans, M. (2007). Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to 2006. *Journal of virology*, 81(18):9932–41.

Siebenga, J. J., Vennema, H., Zheng, D.-P., Vinjé, J., Lee, B. E., Pang, X.-L., Ho, E. C. M., Lim, W., Choudekar, A., Broor, S., Halperin, T., Rasool, N. B. G., Hewitt, J., Greening, G. E., Jin, M., Duan, Z.-J., Lucero, Y., O'Ryan, M., Hoehne, M., Schreier, E., Ratcliff, R. M., White, P. A., Iritani, N., Reuter, G., and Koopmans, M. (2009). Norovirus Illness Is a Global Problem: Emergence and Spread of Norovirus GII.4 Variants, 20012007. *The Journal of Infectious Diseases*, 200(5):802–812.

Simmonds, P., Karakasiliotis, I., Bailey, D., Chaudhry, Y., Evans, D. J., and Goodfellow, I. G. (2008). Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. *Nucleic Acids Research*, 36(8):2530–2546.

Simmons, K., Gambhir, M., Leon, J., and Lopman, B. (2013). Duration of immunity to norovirus gastroenteritis. *Emerging Infectious Diseases*, 19(8):1260–1267.

Singh, B. K., Koromyslova, A., Hefele, L., Gurth, C., and Hansman, G. S. (2016a). Structural Evolution of the Emerging 2014-2015 GII.17 Noroviruses. *J Virol*, 90(5):2710–2715.

Singh, B. K., Leuthold, M. M., and Hansman, G. S. (2015). Human noroviruses' fondness for histo-blood group antigens. *Journal of virology*, 89(4):2024–40.

Singh, B. K., Leuthold, M. M., and Hansman, G. S. (2016b). Structural Constraints on Human Norovirus Binding to Histo-Blood Group Antigens. *MSphere*, 1(2):1–7.

Siqueira, J. A. M., Bandeira, R. D. S., Oliveira, D. D. S., Dos Santos, L. F. P., and Gabbay, Y. B. (2017). Genotype diversity and molecular evolution of noroviruses: A 30-year (1982-2011) comprehensive study with children from Northern Brazil. *PLoS ONE*, 12(6).

Smits, S. L., Rahman, M., Schapendonk, C. M. E., van Leeuwen, M., Faruque, A. S. G., Haagmans, B. L., Endtz, H. P., and Osterhaus, A. D. M. E. (2012). Calicivirus from novel recovirus genogroup in human diarrhea, Bangladesh. *Emerging Infectious Diseases*, 18(7):1192–1195.

Someya, Y., Takeda, N., and Miyamura, T. (2002). Identification of active-site amino acid residues in the Chiba virus 3C-like protease. *Journal of virology*, 76(12):5949–58.

Someya, Y., Takeda, N., and Wakita, T. (2008). Saturation mutagenesis reveals that GLU54 of norovirus 3C-like protease is not essential for the proteolytic activity. *Journal of Biochemistry*, 144(6):771–780.

Sosnovtsev, S. V., Belliot, G., Chang, K.-O., Prikhodko, V. G., Thackray, L. B., Wobus, C. E., Karst, S. M., Virgin, H. W., and Green, K. Y. (2006). Cleavage map and proteolytic processing of the murine norovirus nonstructural polyprotein in infected cells. *Journal of virology*, 80(16):7816–7831.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Subba-Reddy, C. V., Goodfellow, I., and Kao, C. C. (2011). VPg-Primed RNA Synthesis of Norovirus RNA-Dependent RNA Polymerases by Using a Novel Cell-Based Assay. *Journal of Virology*, 85(24):13027–13037.

Subba-Reddy, C. V., Yunus, M. a., Goodfellow, I. G., and Kao, C. C. (2012). Norovirus RNA Synthesis Is Modulated by an Interaction between the Viral RNA-Dependent RNA Polymerase and the Major Capsid Protein, VP1. *Journal of Virology*, 86(18):10138–10149.

Sugieda, M., Nagaoka, H., Kakishima, Y., Ohshita, T., Nakamura, S., and Nakajima, S. (1998). Detection of Norwalk-like virus genes in the caecum contents of pigs. Brief report. *Archives of Virology*, 143(6):1215–1221.

Summa, M., von Bonsdorff, C. H., and Maunula, L. (2012). Pet dogs-A transmission route for human noroviruses? *Journal of Clinical Virology*, 53(3):244–247.

Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., Gray, J. J., Letley, L. H., Rait, G., Tompkins, D. S., and O'Brien, S. J. (2012). Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*, 61(1):69–77.

Tamuri, A. U., Dos Reis, M., Hay, A. J., and Goldstein, R. A. (2009). Identifying changes in selective constraints: Host shifts in influenza. *PLoS Computational Biology*, 5(11).

Taube, S., Jiang, M., and Wobus, C. E. (2010). Glycosphingolipids as receptors for non-enveloped viruses.

Taube, S., Kolawole, A. O., Höhne, M., Wilkinson, J. E., Handley, S. A., Perry, J. W., Thackray, L. B., Akkina, R., and Wobus, C. E. (2013). A mouse model for human norovirus. *mBio*, 4(4).

Teunis, P. F. M., Moe, C. L., Liu, P., Miller, S. E., Lindesmith, L., Baric, R. S., Le Pendu, J., and Calderon, R. L. (2008). Norwalk virus: How infectious is it? *Journal of Medical Virology*, 80(8):1468–1476.

Teunis, P. F. M., Sukhrie, F. H. A., Vennema, H., Bogerman, J., Beersma, M. F. C., and Koopmans, M. P. G. (2015). Shedding of norovirus in symptomatic and asymptomatic infections. *Epidemiology and Infection*, 143(08):1710–1717.

Thackray, L. B., Duan, E., Lazear, H. M., Kambal, A., Schreiber, R. D., Diamond, M. S., and Virgin, H. W. (2012). Critical role for interferon regulatory factor 3 (IRF-3) and IRF-7 in type I interferon-mediated control of murine norovirus replication. *Journal of Virology*, 86(24):13515–13523.

Thorne, L., Arias, A., and Goodfellow, I. (2016). Advances Toward a Norovirus Antiviral: From Classical Inhibitors to Lethal Mutagenesis. *Journal of Infectious Diseases*, 213:S27–S31.

Thorne, L. G. and Goodfellow, I. G. (2014). Norovirus gene expression and replication.

Tomov, V. T., Osborne, L. C., Dolfi, D. V., Sonnenberg, G. F., Monticelli, L. A., Mansfield, K., Virgin, H. W., Artis, D., and Wherry, E. J. (2013). Persistent Enteric Murine Norovirus Infection Is Associated with Functionally Suboptimal Virus-Specific CD8 T Cell Responses. *Journal of Virology*, 87(12):7015–7031.

Tse, H., Lau, S. K. P., Chan, W.-M., Choi, G. K. Y., Woo, P. C. Y., and Yuen, K.-Y. (2012). Complete genome sequences of novel canine noroviruses in Hong Kong. *Journal of virology*, 86(17):9531–2.

Tu, E. T.-V., Bull, R. A., Greening, G. E., Hewitt, J., Lyon, M. J., Marshall, J. A., McIver, C. J., Rawlinson, W. D., and White, P. A. (2008). Epidemics of Gastroenteritis during 2006 Were Associated with the Spread of Norovirus GII.4 Variants 2006a and 2006b. *Clinical Infectious Diseases*, 46(3):413–420.

van Beek, J., Ambert-Balay, K., Botteldoorn, N., Eden, J. S., Fonager, J., Hewitt, J., Iritani, N., Kroneman, A., Vennema, H., Vinj??, J., White, P. A., Koopmans, M., and NoroNet (2013). Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. *Euro surveillance : bulletin europ??en sur les maladies transmissibles = European communicable disease bulletin*, 18(1):8–9.

Van Beek, J., De Graaf, M., Xia, M., Jiang, X., Vinjé, J., Beersma, M., De Bruin, E., Van De Vijver, D., Holwerda, M., Van Houten, M., Buisman, A. M., Van Binnendijk, R., Osterhaus, A. D., Van Der Klis, F., Vennema, H., and Koopmans, M. P. (2016). Comparison of norovirus genogroup I, II and IV seroprevalence among children in the Netherlands, 1963, 1983 and 2006. *Journal of General Virology*, 97(9):2255–2264.

Vashist, S., Bailey, D., Putics, A., and Goodfellow, I. (2009). Model systems for the study of human norovirus biology. *Future Virology*, 4(4):353–367.

Vega, E., Barclay, L., Gregoricus, N., Shirley, S. H., Lee, D., and Vinjé, J. (2014a). Genotypic and epidemiologic trends of norovirus outbreaks in the united states, 2009 to 2013. *Journal of Clinical Microbiology*, 52(1):147–155.

Vega, E., Barclay, L., Gregoricus, N., Williams, K., Lee, D., and Vinjé, J. (2011). Novel surveillance network for norovirus gastroenteritis outbreaks, United States. *Emerging Infectious Diseases*, 17(8):1389–1395.

Vega, E., Donaldson, E., Huynh, J., Barclay, L., Lopman, B., Baric, R., Chen, L. F., and Vinjé, J. (2014b). RNA populations in immunocompromised patients as reservoirs for novel norovirus variants. *Journal of virology*, 88(24):14184–96.

Verhoef, L., Hewitt, J., Barclay, L., Ahmed, S. M., Lake, R., Hall, A. J., Lopman, B., Kroneman, A., Vennema, H., Vinjé, J., and Koopmans, M. (2015). Norovirus genotype

profiles associated with foodborne transmission, 1999???2012. *Emerging Infectious Diseases*, 21(4):592–599.

Vinje, J. (2015). Advances in Laboratory Methods for Detection and Typing of Norovirus. *J Clin Microbiol*, 53(2):373–381.

Vongpunsawad, S., Venkataram Prasad, B. V., and Estes, M. K. (2013). Norwalk Virus Minor Capsid Protein VP2 Associates within the VP1 Shell Domain. *Journal of virology*, 87(9):4818–25.

Wang, Q. H., Myung, G. H., Cheetham, S., Souza, M., Funk, J. A., and Saif, L. J. (2005). Porcine noroviruses related to human noroviruses. *Emerging Infectious Diseases*, 11(12):1874–1881.

Ward, J. M., Wobus, C. E., Thackray, L. B., Erexson, C. R., Faucette, L. J., Belliot, G., Barron, E. L., Sosnovtsev, S. V., and Green, K. Y. (2006). Pathology of immunodeficient mice with naturally occurring murine norovirus infection. *Toxicologic Pathology*, 34(6):708–715.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J., and Roderick, P. J. (1999). Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. *BMJ (Clinical research ed.)*, 318(7190):1046–50.

White, P. A. (2014). Evolution of norovirus.

Widdowson, M.-A., Monroe, S. S., and Glass, R. I. (2005). Are Noroviruses Emerging? *Emerging Infectious Diseases*, 11(5):735–737.

Wobus, C. E., Karst, S. M., Thackray, L. B., Chang, K. O., Sosnovtsev, S. V., Belliot, G., Krug, A., Mackenzie, J. M., Green, K. Y., and Virgin IV, H. W. (2004). Replication of Norovirus in cell culture reveals a tropism for dendritic cells and macrophages. *PLoS Biology*, 2(12).

Wobus, C. E., Thackray, L. B., and Virgin, H. W. (2006). Murine Norovirus: a Model System To Study Norovirus Biology and Pathogenesis. *Journal of Virology*, 80(11):5104–5112.

Wolf, S., Williamson, W., Hewitt, J., Lin, S., Rivera-Aban, M., Ball, A., Scholes, P., Savill, M., and Greening, G. E. (2009). Molecular detection of norovirus in sheep and pigs in New Zealand farms. *Veterinary Microbiology*, 133(1-2):184–189.

Wong, T. H. N., Dearlove, B. L., Hedge, J., Giess, A. P., Piazza, P., Trebes, A., Paul, J., Smit, E., Smith, E. G., Sutton, J. K., Wilcox, M. H., Dingle, K. E., Peto, T. E. a., Crook, D. W., Wilson, D. J., and Wyllie, D. H. (2013). Whole genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England. *Virology journal*, 10:335.

Wright, P. J., Gunesekere, I. C., Doultree, J. C., and Marshall, J. A. (1998). Small round-structued (Norwalk-like) viruses and classical human caliciviruses in Southeastern Australia, 1980-1996. *Journal of Medical Virology*, 55(4):312–320.

Wu, Z., Yang, L., Ren, X., He, G., Zhang, J., Yang, J., Qian, Z., Dong, J., Sun, L., Zhu, Y., Du, J., Yang, F., Zhang, S., and Jin, Q. (2016). Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *The ISME Journal*, 10(3):609–620.

Xi, J. N., Graham, D. Y., Wang, K. N., and Estes, M. K. (1990). Norwalk virus genome cloning and characterization. *Science*, 250(4987):1580–1583.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Zahorsky, J. (1929). Hyperemesis Hiemis or the winter vomiting disease. *Archives of Pediatrics & Adolescent Medicine*, 46:391–395.

Zakikhany, K., Allen, D. J., Brown, D., and Iturriza-Gómara, M. (2012). Molecular evolution of GII-4 norovirus strains. *PLoS ONE*, 7(7).

Zamyatkin, D. F., Parra, F., Alonso, J. M. M., Harki, D. a., Peterson, B. R., Grochulski, P., and Ng, K. K.-S. (2008). Structural insights into mechanisms of catalysis and inhibition

in Norwalk virus polymerase. *The Journal of biological chemistry*, 283(12):7705–7712.

Zeitler, C. E., Estes, M. K., and Venkataram Prasad, B. V. (2006). X-ray crystallographic structure of the Norwalk virus protease at 1.5-A resolution. *Journal of virology*, 80(10):5050–8.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9(1):40.

Zheng, D. P., Ando, T., Fankhauser, R. L., Beard, R. S., Glass, R. I., and Monroe, S. S. (2006). Norovirus classification and proposed strain nomenclature. *Virology*, 346(2):312–323.

Zhu, S., Regev, D., Watanabe, M., Hickman, D., Moussatche, N., Jesus, D. M., Kahan, S. M., Napthine, S., Brierley, I., Hunter, R. N., Devabhaktuni, D., Jones, M. K., and Karst, S. M. (2013). Identification of Immune and Viral Correlates of Norovirus Protective Immunity through Comparative Study of Intra-Cluster Norovirus Strains. *PLoS Pathogens*, 9(9).

Appendix A

Supplementary figures and tables

A.1 Chapter 2 supplementary figures and tables

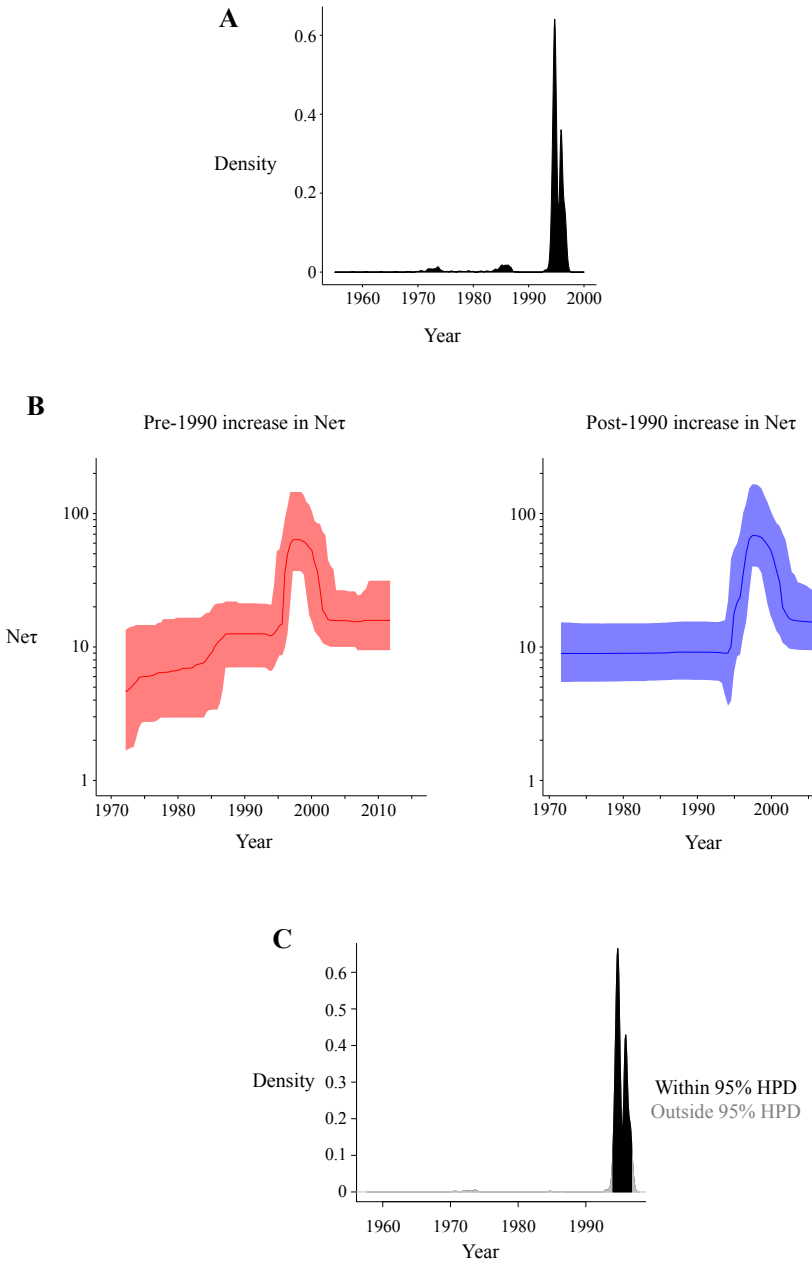


Figure S2.1: The distribution of the time of GII.4 pandemic onset. (**A**) The distribution of the time at which the Bayesian skyline plot first exhibits an increase in Neτ of more the 100% relative to baseline. We calculated this time in each sampled step in the MCMC chain. While most samples support the first increase occurring in the mid-1990s, a small number of samples support the first increase occurring prior to 1990. (**B**) Bayesian skyline plots for MCMC samples supporting the first increase in Neτ prior to 1990 (red) and post-1990 (blue). The samples supporting an early increase in Neτ have a small Neτ within the earliest time slice and exhibit a small increase in Neτ prior to a large increase in the mid-1990s, coinciding with the first increase in Neτ in the post-1990 increase samples. We therefore defined the time of pandemic onset within the pre-1990 increase samples as the time at which Neτ increased by more than 400%. (**C**) The distribution of the time of increase in GII.4 frequency, calculated as the time at which Neτ increased by more than 100% relative to baseline in the post-1990 increase samples or by more than 400% in the pre-1990 increase samples. The black shaded region shows the 95% HPD of the time of pandemic emergence.

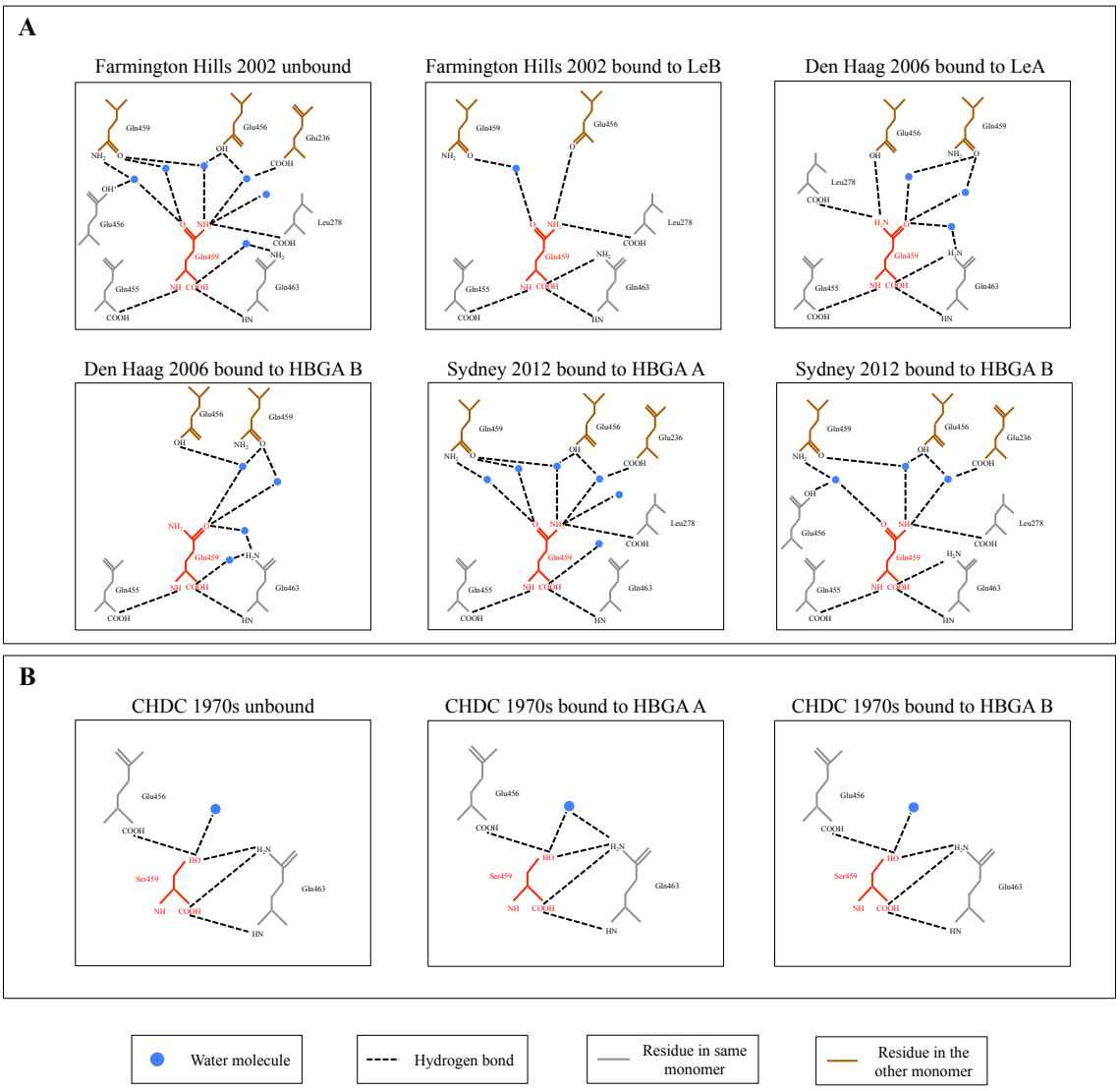


Figure S2.2: Comparison of the interaction network at site 459 in solved crystal structures. The residue at site 459 is shown in red in each case and residues that hydrogen bond with site 459 either directly or through a single intermediate water molecule are shown in grey if they are in the same monomer or brown if they are in the other monomer. Water molecules are shown as blue circles and each hydrogen bond is shown as a dashed black line. (**A**) The hydrogen bond network formed by glutamine 459 is shown for six solved P domain structures from the pandemic GII.4 clade: Farmington Hills 2002 unbound (PDB identifier 4OOV), Farmington Hills 2002 bound to Lewis B (4OPS), Den Haag 2006 bound to Lewis A (4WZL), Den Haag 2006 bound to HBGA type B (4X06), Sydney 2012 bound to HBGA type A (4WZT) and Sydney 2012 bound to HBGA type B (4OP7). The structures from the same strain have the same viral sequence. (**B**) The hydrogen bond network formed by serine 459 is shown for three solved P domain structures from a CHDC 1970s virus (pre-pandemic): unbound (PDB identifier 5IYN), bound to HBGA type A (5IYP) and bound to HBGA type B (5IYQ). These three structures have the same viral sequence. While there are differences in the exact hydrogen bond network formed within different pandemic GII.4 structures and within different CHDC 1970s structures, the network is consistently more extensive in the pandemic GII.4 structures compared with the pre-pandemic GII.4 structures.

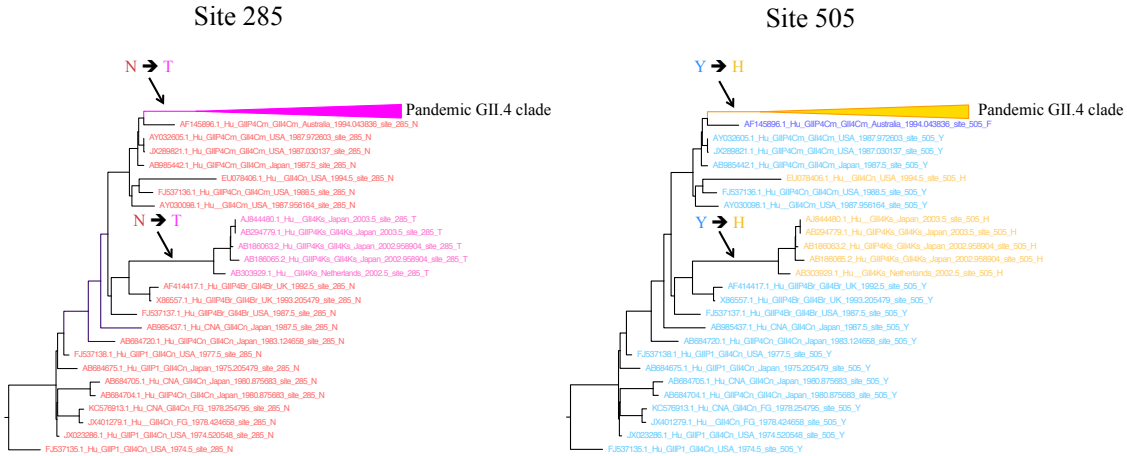
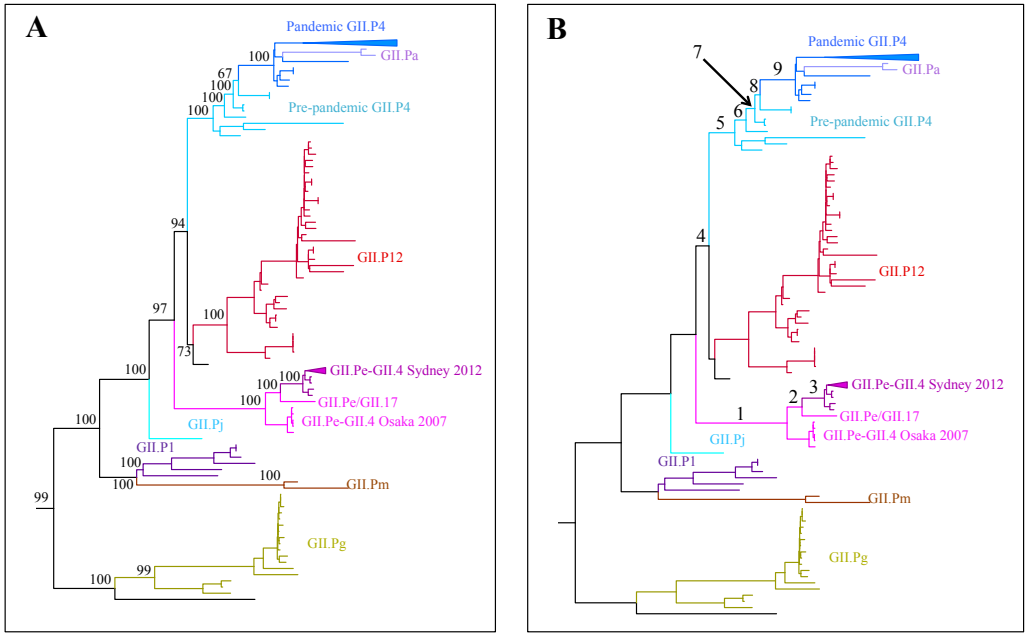


Figure S2.3: Convergent substitutions at sites 285 and 505. Coloured trees of the early GII.4 lineage for sites 285 and 505. Each site exhibits a convergent substitution leading to the Kaiso 2003 strain and leading to the pandemic GII.4 clade.



C

Branch	ORF1 nonsynonymous substitutions
1	T20I, S31N, R50K, Q54P, P79L, V85A, S96T, D102E, L130I, V149I, A261T, R278K, S288T, A463S, D618E, M650S, V710I, R732K, R738K, E742D, I750V, R754K, S768A, N773S, R778K, V830A, D837N, D983E, R1004K, Q1045K, E1156G, K1268R, R1420K, A1442D, K1480Q, N1484S, N1515D, N1575D
2	T699A, S751A, T883S, S1270N
3	A13V, L29F, N118S, V750I, G760N, T850M, S852N, P857L, K927R, N1193S, I1364V, K1425R, S1616F
4	K511R, G760S, S762C, S852N, A1296T
5	A13V, P79L, T104A, R189K, M650A, R729K, A851T
6	K496R, A650T, N1006S, N1193S, M1314I
7	None
8	R190K, R1004K
9	L79I, D102E, I342V, V491L, L508V, I525V, A586V, D618E, P642S, Y693F, T699A, R754K, T783S, T850A, V867I, Q1045K, A1091V, Q1291E, I1314M, V1404I, P1456S, I1463V, M1474V

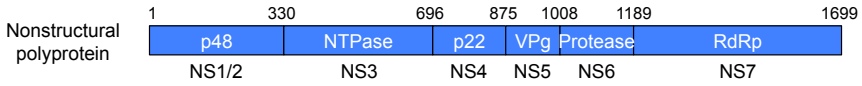


Figure S2.4: Nonsynonymous substitutions close to the GII.P4 and GII.Pe genotypes. (**A**) Maximum likelihood phylogeny showing the evolutionary relationship between GII.P4, GII.Pe and closely related genotypes. Branches are coloured by strain with the colour matching that of the strain label. Bootstrap supports are shown at trunk nodes. Part of the pandemic GII.P4 clade has been collapsed for clarity. This clade was extracted from a nucleotide maximum likelihood tree reconstructed on all GII and GIV ORF1 sequences. (**B**) Ancestral reconstruction was carried out to determine the substitutions occurring along each branch within ORF1. Branches 1-9 lead to the pandemic ORF1 clades following their divergence. Therefore the independent evolution leading to these clades following their divergence occurred along these branches. (**C**) The nonsynonymous substitutions inferred to occur along each of the branches 1-9 are shown. Substitutions in red occurred convergently leading to the pandemic GII.P4 clade (along any of the branches 4-9) and the pandemic GII.Pe clade (along any of the branches 1-3). Substitutions in blue represent sites that changed leading to both pandemic clades but the residue change was different in each case. Note that while there are pairs of sites that are located close together in the ORF1 sequence that change leading to each pandemic clade, the residues at these sites are different in each case and there is therefore no reason to expect that they would result in similar phenotypic changes. The nonstructural polyprotein is shown with the residues delimiting each individual protein labelled (Belliot et al., 2003)

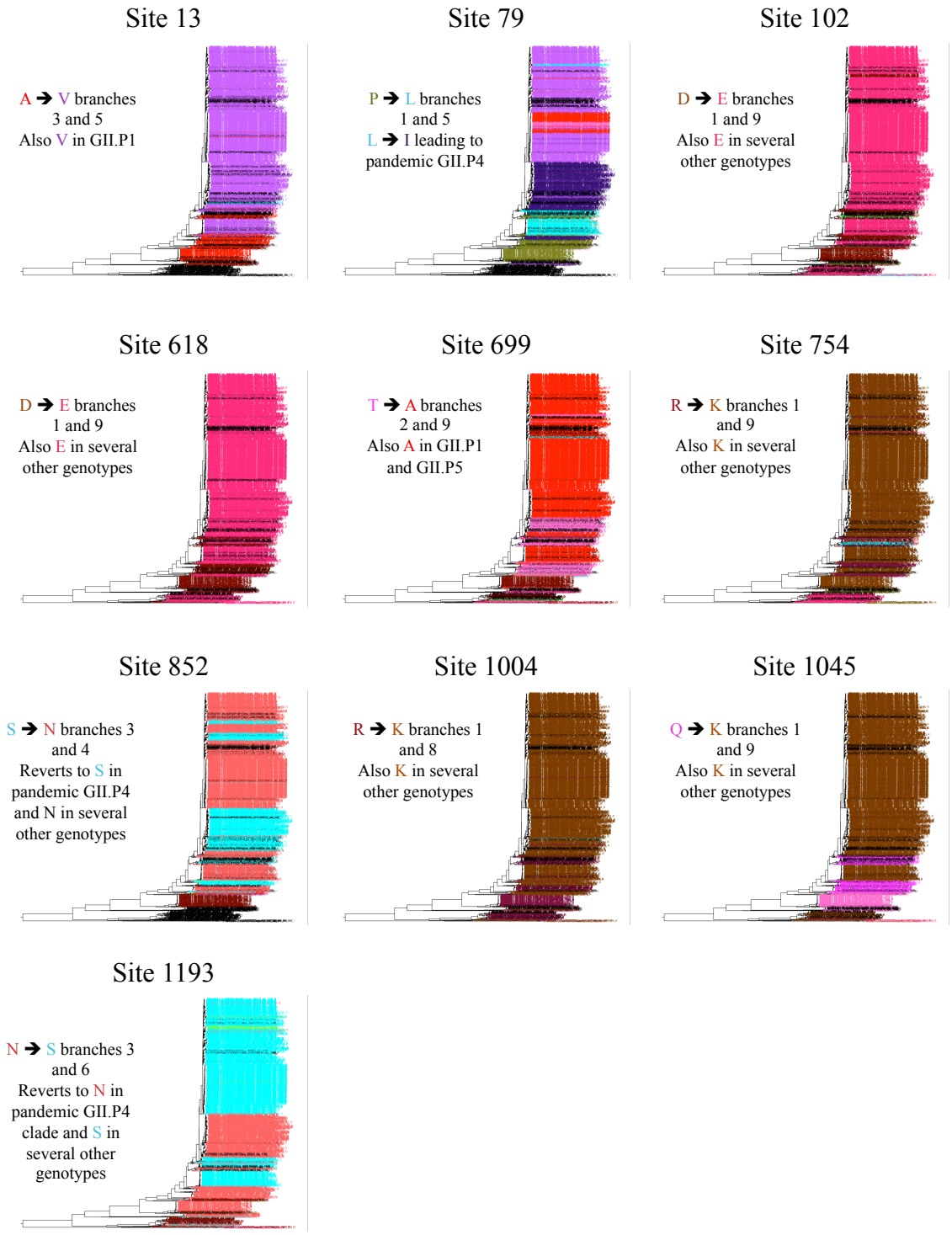


Figure S2.5: Convergent substitutions leading to the GII.P4 and GII.Pe genotypes. A coloured tree is shown for each of the sites that exhibit convergent substitution leading to the pandemic GII.P4 and GII.Pe clades after their divergence (red substitutions in figure S2.4 panel C). For each site, the substitution is listed along with the branches on which the substitution occurred. We suggest that none of these substitutions were essential for pandemic emergence due to either the residue also being present in other, non-pandemic, genotypes, the substitution reverting in one or more pandemic strains within the pandemic GII.P4 clade or, in the case of site 79, the site undergoing a further substitution leading to the pandemic GII.P4 clade. The sites where the residue present in both pandemic clades is also present in other genotypes are typically highly variable between two or more residues, with multiple substitutions occurring between these residues. The tree shown is the nucleotide maximum likelihood tree reconstructed on all available ORF1 sequences from the GII and GIV genogroups.

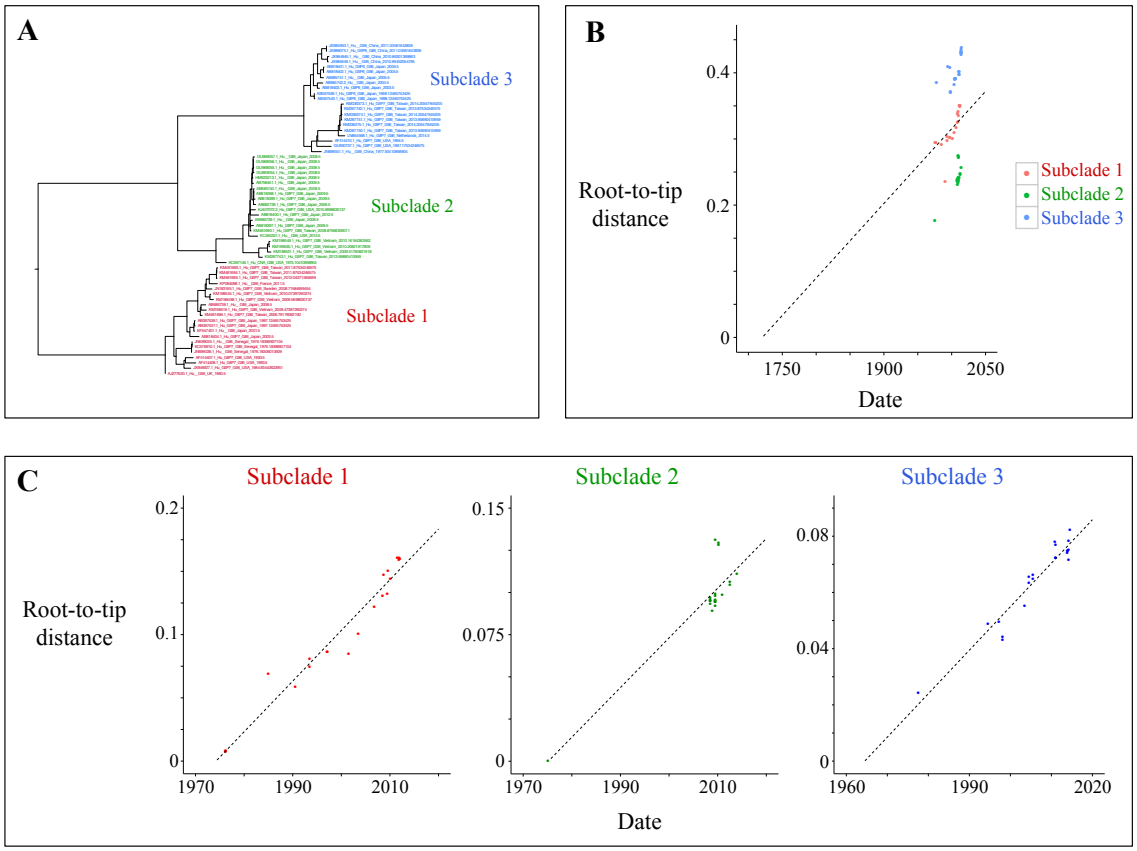


Figure S2.6: Subclades within GII.6 exhibit accumulation of nucleotide change through time. **(A)** Maximum likelihood tree of the GII.6 capsid genotype. The tree exhibits three well diverged subclades, which we named subclade 1 (red), subclade 2 (green) and subclade 3 (blue). **(B)** The correlation between root-to-tip distance and collection date within the GII.6 capsid genotype nucleotide maximum likelihood tree. The points are coloured by subclade. While there is no correlation when including all subclades, each subclade appears to exhibit a correlation. **(C)** We reconstructed a nucleotide maximum likelihood tree on each GII.6 subclade independently. Plotted here is the root-to-tip distance versus collection date for each subclade.

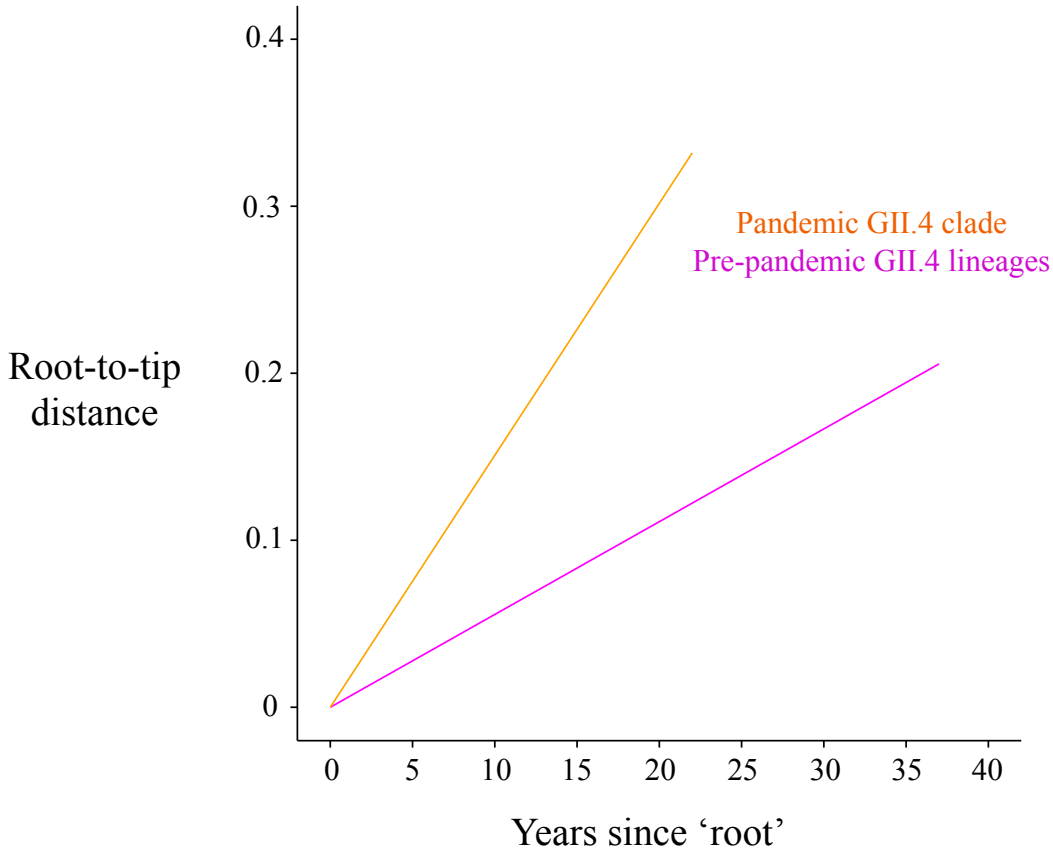


Figure S2.7: Comparison of the magnitude of amino acid change within the pandemic GII.4 clade and pre-pandemic GII.4 lineages. We extracted the pandemic GII.4 clade and the pre-pandemic GII.4 lineages from the phylogenetic tree in Figure 2.1 and optimised branch lengths independently using the respective amino acid alignment and the WAG substitution matrix. We then plotted the accumulation of amino acid change versus collection date for each sequence within each clade and calculated the best fit regression line. Both clades exhibit a correlation between amino acid root-to-tip distance and collection date. Plotted here are the regression lines normalised so the root of the respective phylogeny occurred at time zero. There is a greater magnitude of amino acid change in the pandemic GII.4 clade compared with the pre-pandemic GII.4 lineages.

Genotype	Putative breakpoint	Number of samples removed	Accession numbers of removed samples
GII.12	846	2	EF547403.1, KC464498.1
GII.14	273	1	GU017903.2
GII.21	1014	6	AY675554.1, EU019230.2, GQ856468.1 KJ196284.1, JN899245.1, KU687014.1

Table S2.1: Putative recombinant samples removed from capsid genotype datasets. We screened for recombination within each capsid genotype dataset using SBP. Where a significant breakpoint was identified, we reconstructed a maximum likelihood tree on either side of the putative breakpoint and classified sequences that clustered differently on each side of the breakpoint with strong bootstrap support as being putatively recombinant. These sequences were removed from further analyses. Putative recombination was identified within the GII.12, GII.14 and GII.21 genotypes. The breakpoint here is the most likely breakpoint position within the capsid identified by SBP and may not be the true recombination breakpoint. Removal of the sequences in this table resulted in loss of the recombination signal.

A.2 Chapter 3 supplementary figures and tables

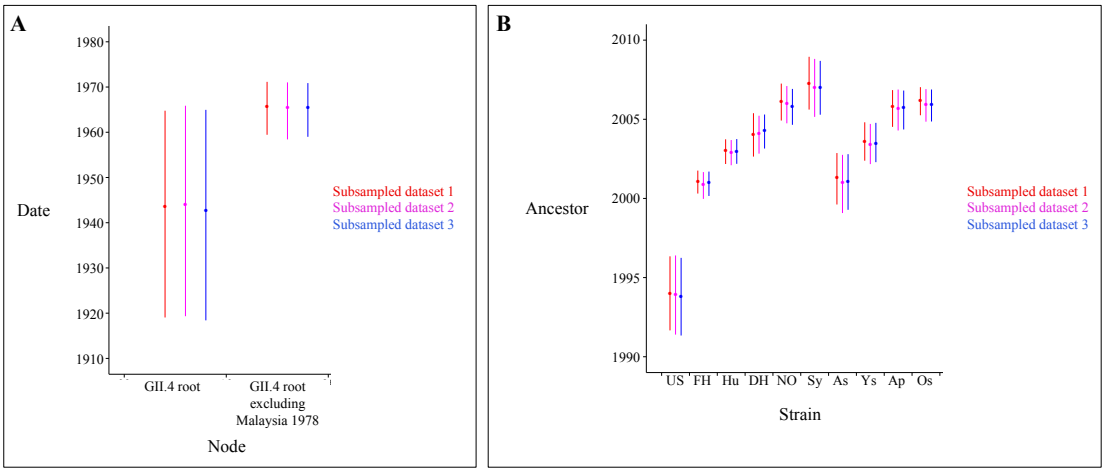


Figure S3.1: Comparison of node and ancestor dates between subsampled datasets. (**A**) The mean (circle) and 95% HPD is shown for the date of the root node of the GII.4 genotype and the node one downstream of the root node following the divergence of a sequence collected in Malaysia in 1978. The date of the GII.4 root excluding the Malaysia 1978 sequence is comparable to that obtained in a previous study (Bok et al., 2009) where the Malaysia 1978 sequence was not included. (**B**) As in panel A, but the date of the common ancestor of each GII.4 strain is shown. US - US95/96, FH - Farmington Hills 2002, Hu - Hunter 2004, DH - Den Haag 2006, NO - New Orleans 2009, Sy - Sydney 2012, As - Asia 2003, Ys - Yerseke 2006, Ap - Apeldoorn 2007, Os - Osaka 2007.

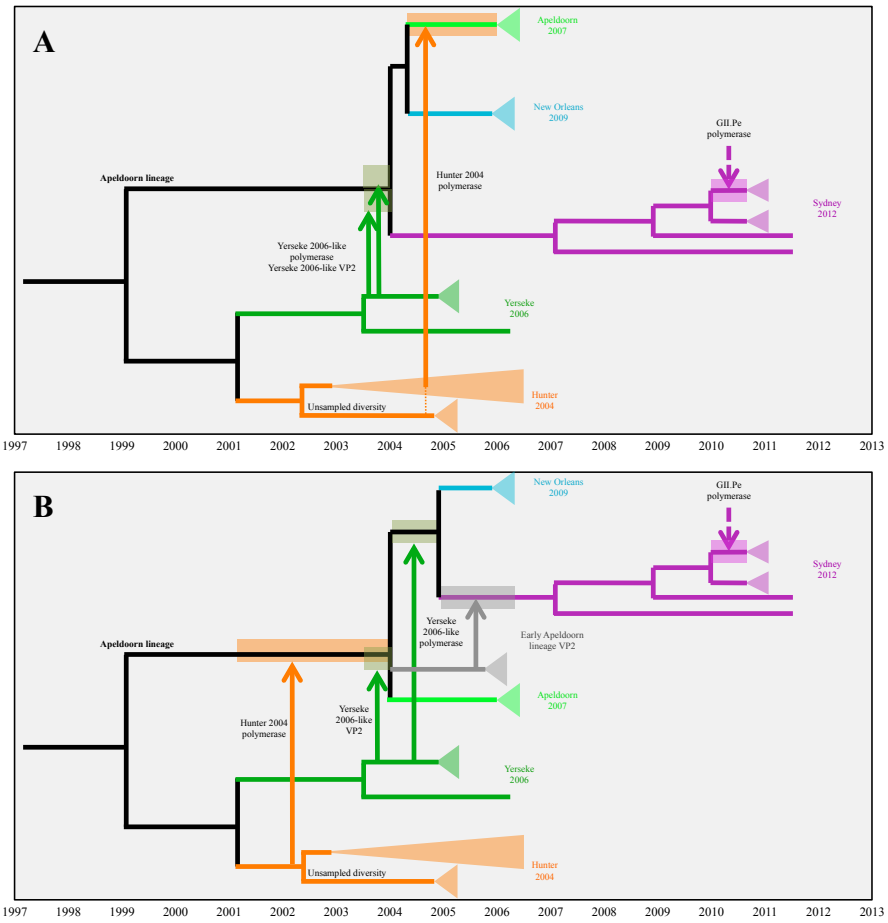


Figure S3.2: Recombination events in the Apeldoorn lineage. Shown are two hypothetical scenarios for the acquisition of RdRp and VP2 genomic regions by the Apeldoorn lineage capsid (consisting of Apeldoorn 2007, New Orleans 2009 and Sydney 2012) that are consistent with the tree topologies and divergence dates in Figure 3.1. The tree backbones are from the capsid tree in Figure 3.1 in which the relationship between Apeldoorn 2007, New Orleans 2009 and Sydney 2012 is uncertain. Therefore each scenario depicts a hypothetical relationship between these strains that fits that scenario. Arrows represent a recombination event in which a RdRp or VP2 genomic region is obtained, with the arrow pointing from donor strain to recipient strain. Triangles at the end of branches represent the remaining lineages within the strain. **(A)** This scenario involves the acquisition of a Yerseke 2006-like VP2 prior to the divergence of the strains within the lineage. The relationship between Apeldoorn 2007, New Orleans 2009 and Sydney 2012 is the same as the well-supported relationship within the VP2 tree, with Sydney 2012 diverging first. The Yerseke 2006 RdRp is acquired prior to divergence of strains within the Apeldoorn lineage and Apeldoorn 2007 acquires a Hunter 2004-like RdRp after its divergence from New Orleans 2009. **(B)** This scenario involves acquisition of the Yerseke 2006 VP2 and a Hunter 2004-like RdRp prior to divergence of the strains within the Apeldoorn lineage. Here, Apeldoorn 2007 branches first and the Yerseke 2006-like RdRp is acquired leading to the common ancestor of New Orleans 2009 and Sydney 2012. Sydney 2012 acquires its VP2 region from an early Apeldoorn lineage virus that has persisted, explaining the divergence of Sydney 2012 in the VP2 tree. In both scenarios, Sydney 2012 acquires the GII.Pe RdRp after divergence from Apeldoorn 2007 and New Orleans 2009. Scenario **B** requires one more recombination event than scenario **A** and requires the persistence of an unsampled early Apeldoorn lineage.

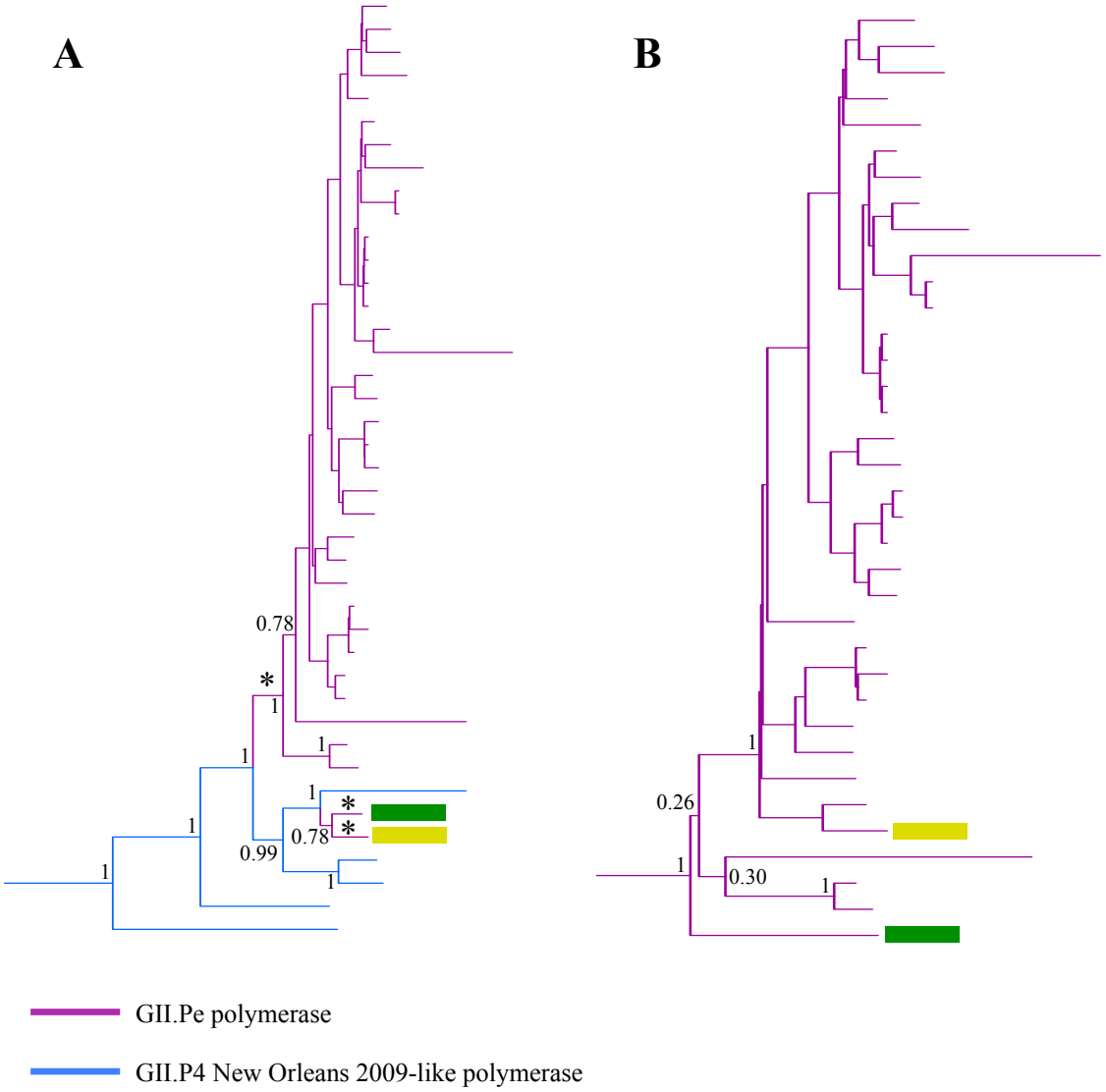


Figure S3.3: Acquisition of the GII.Pe RdRp by the Sydney 2012 capsid. **(A)** The Sydney 2012 clade in the capsid MCC tree in Figure 3.1 is shown. Tip branches are coloured by the RdRp found with that sequence; blue - GII.4 New Orleans 2009-like RdRp, purple - GII.Pe RdRp. Acquisition of the GII.Pe RdRp by recombination has been inferred to occur along the starred branches. **(B)** The GII.Pe-Sydney 2012 clade in the RdRp MCC tree in Figure 3.1 is shown. While the sequences marked with green and yellow rectangles each have the GII.Pe RdRp and are monophyletic in the capsid tree, they are not monophyletic in the RdRp tree and therefore must have been acquired in separate recombination events. As the GII.Pe RdRp marked with a yellow rectangle clusters within the GII.Pe-Sydney 2012 clade in the RdRp tree, it was likely acquired from a virus with a Sydney 2012 capsid. Posterior supports are shown at key nodes.

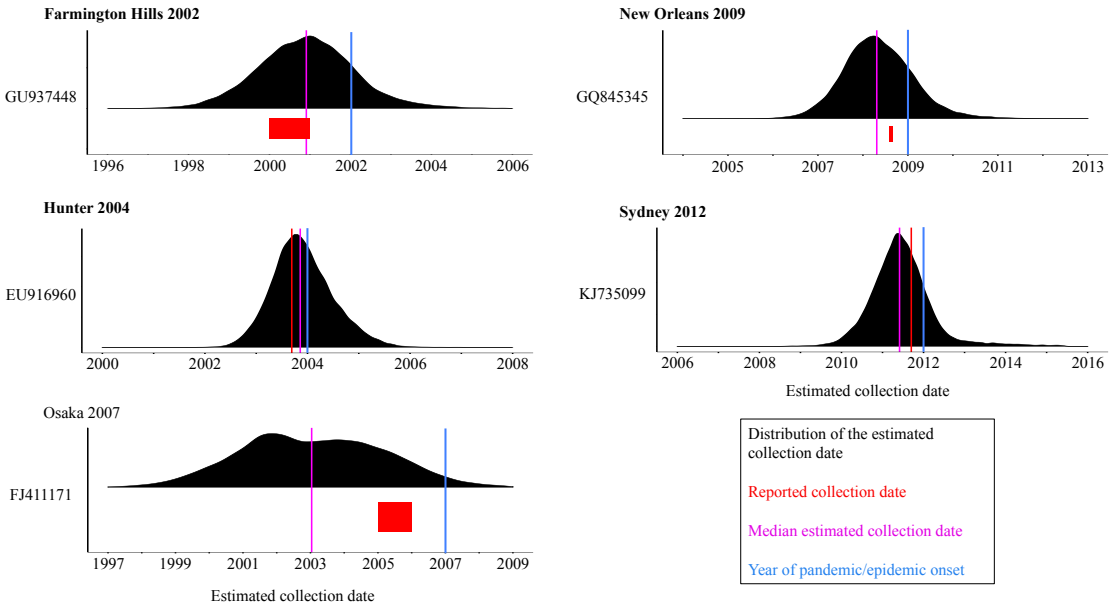


Figure S3.4: Example collection date distributions for pre-pandemic and pre-epidemic sequences. We identified 50 putative pre-pandemic and pre-epidemic GII.4 sequences and estimated the collection date of each sequence using BEAST. The estimated collection date overlapping with the reported collection date was taken as evidence to support, but not confirm, the reported collection date, suggesting that these sequences are true pre-pandemic/pre-epidemic sequences. Shown here is the posterior distribution of the estimated collection date for one sequence from each strain with a pre-pandemic or pre-epidemic sequence. The median of this distribution is shown by the magenta vertical line. The reported collection date is shown by the red line/red area; where the collection date was reported to the day this is shown by a vertical red line, while if the collection date was reported to the nearest month or year, the red area shows this time span. The red reported collection date overlapping with the black distribution of the estimated collection date was taken as evidence to support the reported collection date. The start of the year in which the respective pandemic or epidemic strain emerged is shown by the blue vertical line. The accession number of each sequence is shown next to the respective distribution and details of each sequence are summarised in Table 3.6.

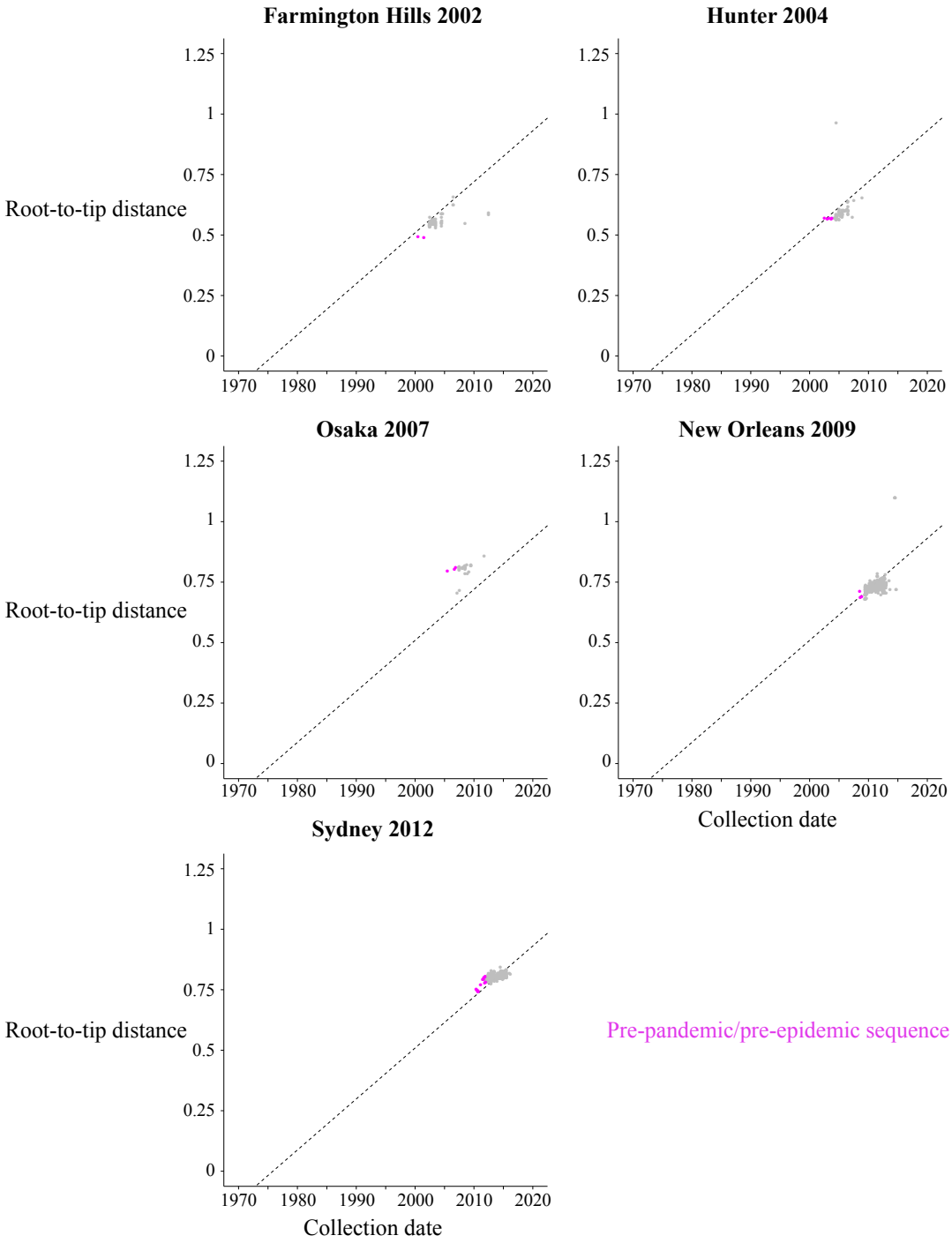


Figure S3.5: Accumulation of nucleotide change by putative pre-pandemic/pre-epidemic sequences. We reconstructed a nucleotide maximum likelihood tree of the P2 domain of all available GII.4 sequences and rooted the tree to minimise the heuristic residual mean squared score. Plotted for each strain with a putative pre-pandemic or pre-epidemic sequence is the root-to-tip distance versus collection date for each sequence within the strain. The putative pre-pandemic/pre-epidemic sequences where the collection date estimated with BEAST overlaps with the reported collection date are coloured magenta, other sequences from the strain are coloured grey. Each of the putative pre-pandemic/pre-epidemic sequences exhibits close to the expected level of nucleotide change for their reported collection date, given the accumulation of nucleotide change exhibited by other sequences within the same strain. The trendline shows the general rate of accumulation of change within the GII.4 genotype, as calculated from the maximum likelihood phylogenetic tree.

Most likely recombination breakpoint	GII.4 strain 5' to breakpoint	GII.4 strain 3' to breakpoint	Number of sequences	Accession numbers
537	Den Haag 2006	New Orleans 2009	12	KF712501.1, KF429790.1, KF712491.1, KF429762.1, KF712498.1, KF712505.1, KF712495.1, KF429785.1, KF429776.1, JX459900.1, KF712502.1, AB933738.1
537	Den Haag 2006	Apeldoorn 2007	1	AB541362.1
537	New Orleans 2009	Apeldoorn 2007	1	JX448566.1
314	Den Haag 2006	Apeldoorn 2007	1	KF712492.1
314	New Orleans 2009	Den Haag 2006	3	KF196287.1, AB933682.1, AB447434.1
314	Unclear	Osaka 2007	1	GQ845368.2

Table S3.1: Summary of putative recombination events in the GII.4 capsid. Putative recombination events were identified using SBP. We identified the most likely recombination breakpoint as the position with the greatest gain in AIC when splitting the alignment at that position compared with the null model. As SBP can only examine recombination at variable positions, this may not be the true recombination breakpoint. The number of sequences that cluster together but in different clades on either side of the breakpoint is shown. Removal of the sequences in this table resulted in loss of the recombination signal.

	RdRp ancestor			Capsid ancestor			VP2 ancestor		
GII.4 strain	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
US95/96	1993 (1991,1995)	1993 (1991,1996)	1994 (1992,1996)	1994 (1991,1996)	1993 (1991,1996)	1993 (1991,1996)	1994 (1993,1996)	1994 (1993,1996)	1994 (1993,1996)
Farmington Hills 2002	2000 (1999,2001)	2000 (1999,2001)	2000 (1999,2001)	2001 (2000,2001)	2000 (1999,2001)	2001 (2000,2001)	1999 (1998,2000)	1999 (1998,2000)	2000 (1998,2000)
Asia 2003	2002 (2001,2003)	2002 (2001,2003)	2002 (2001,2003)	2001 (1999,2002)	2001 (1999,2002)	2001 (1999,2002)	2001 (2000,2002)	2001 (1999,2002)	2001 (2000,2002)
Hunter 2004	2002 (2001,2003)	2002 (2001,2003)	2002 (2001,2003)	2003 (2002,2003)	2002 (2002,2003)	2002 (2002,2003)	2002 (2001,2003)	2002 (2001,2003)	2002 (2001,2003)
Yerseke 2006	2002 (2001,2003)	2002 (2001,2003)	2002 (2001,2003)	2003 (2002,2004)	2003 (2002,2004)	2003 (2002,2004)	2002 (2001,2003)	2002 (2001,2003)	2002 (2002,2004)
Den Haag 2006	2004 (2003,2005)	2004 (2003,2005)	2003 (2002,2004)	2004 (2002,2005)	2004 (2002,2005)	2004 (2003,2005)	2002 (2001,2003)	2002 (2001,2004)	2003 (2002,2004)
Osaka 2007	2006 (2005,2007)	2006 (2005,2007)	2006 (2005,2006)	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)	2005 (2003,2006)	2005 (2004,2006)
Apeldoorn 2007	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)	2006 (2005,2007)	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)	2005 (2004,2006)
New Orleans 2009	2005 (2003,2006)	2005 (2004,2006)	2005 (2003,2006)	2006 (2004,2007)	2005 (2004,2007)	2005 (2004,2006)	2006 (2005,2006)	2005 (2004,2006)	2005 (2005,2006)
Sydney 2012	2009 (2008,2010)	2009 (2008,2010)	2009 (2008,2010)	2007 (2005,2008)	2007 (2005,2008)	2006 (2005,2008)	2005 (2003,2007)	2005 (2003,2007)	2005 (2003,2007)

Table S3.2: Comparison of strain ancestor dates in each subsampled dataset. Samples 1, 2 and 3 contain a different subset of 41 sequences from Den Haag 2006 and 41 sequences from New Orleans 2009. The Sydney 2012 RdRp ancestor date is the common ancestor of the GII.Pe RdRps associated with the Sydney 2012 capsid.

A.3 Chapter 4 supplementary figures

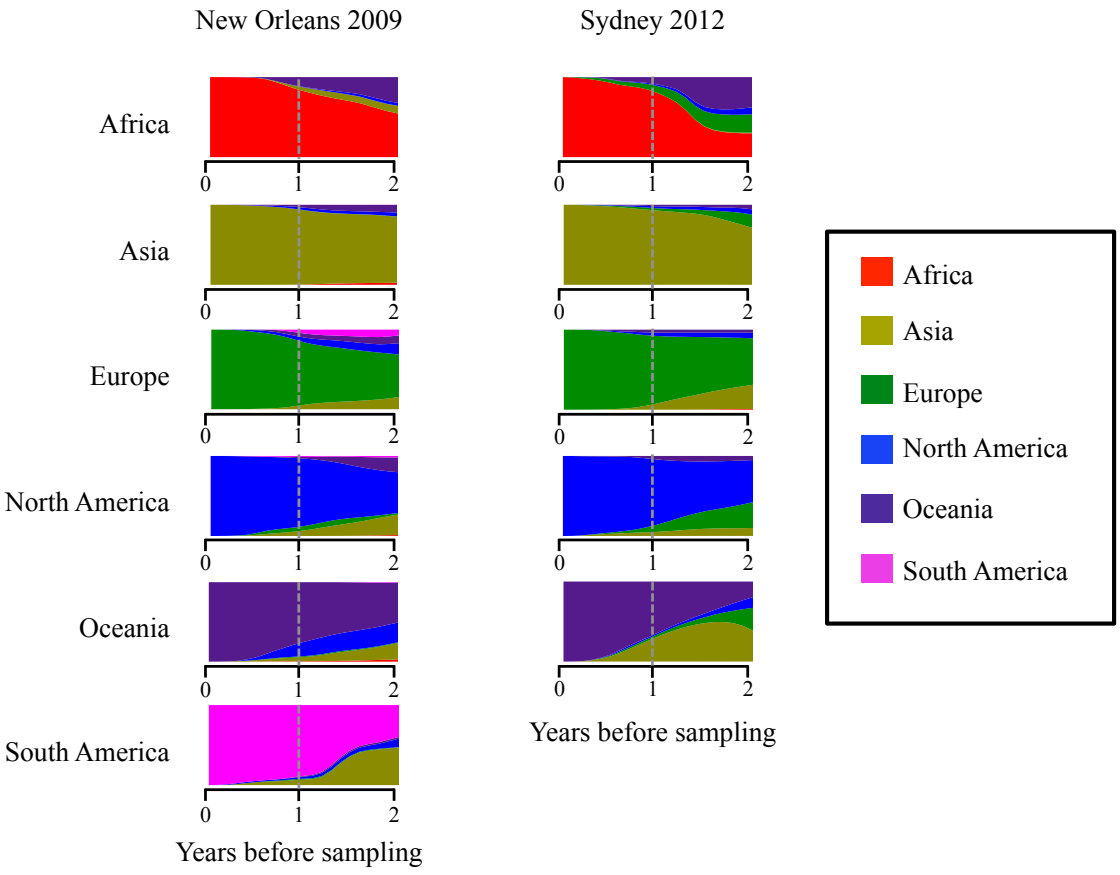


Figure S4.1: The majority of viral lineages were imported into the continent of collection more than a year prior to sampling. We collected each phylogeny tip belonging to a continent and traced the phylogeographic history of that lineage back in time. Shown is the average ancestry of viruses collected within each continent, averaged across all tips from the region within a tree and across the posterior distribution of trees. The average ancestry is shown as a stacked area plot. As an example, the top left panel shows the average ancestry of New Orleans 2009 viruses collected in Africa. At time 0, all viruses are by definition within the continent in which they were collected. By one year prior to sampling, the majority of ancestors were already in Africa, while a small number of ancestors were yet to migrate into Africa and at that time were in Oceania or Asia. The grey vertical dashed lines represent one year prior to sampling, at which time approximately 86% and 85% of New Orleans 2009 and Sydney 2012 lineages, respectively, had been imported into their continent of sampling.

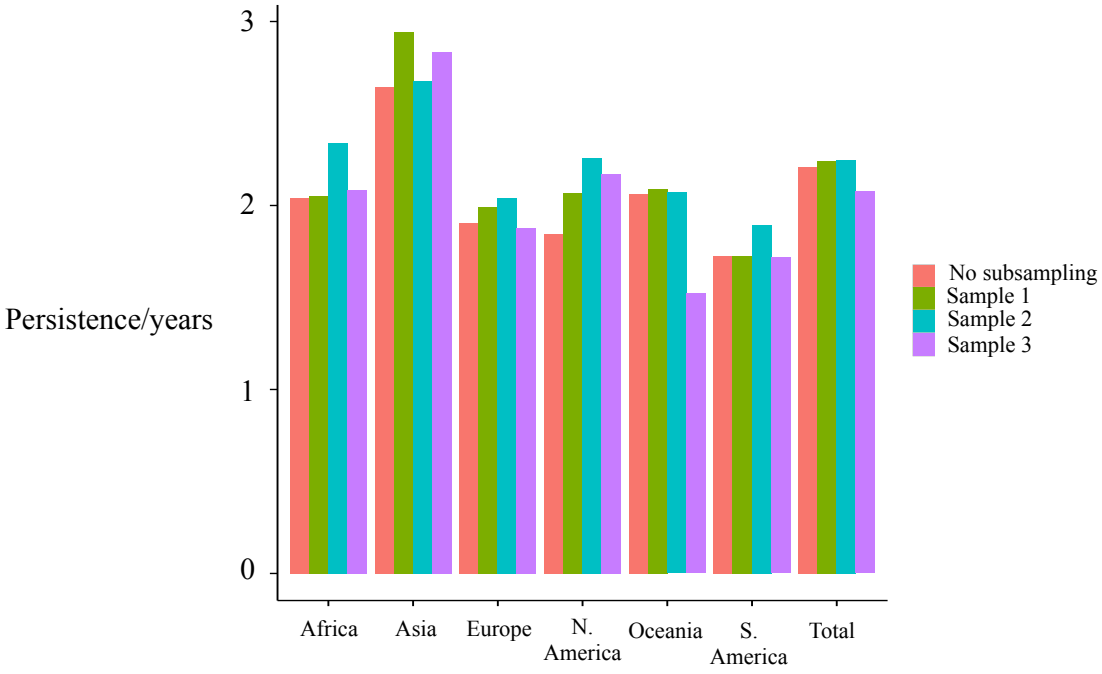


Figure S4.2: Subsampling does not alter estimates of sample persistence. The mean duration of persistence on each continent is plotted for each subsampled New Orleans 2009 dataset and a New Orleans 2009 dataset without subsampling.

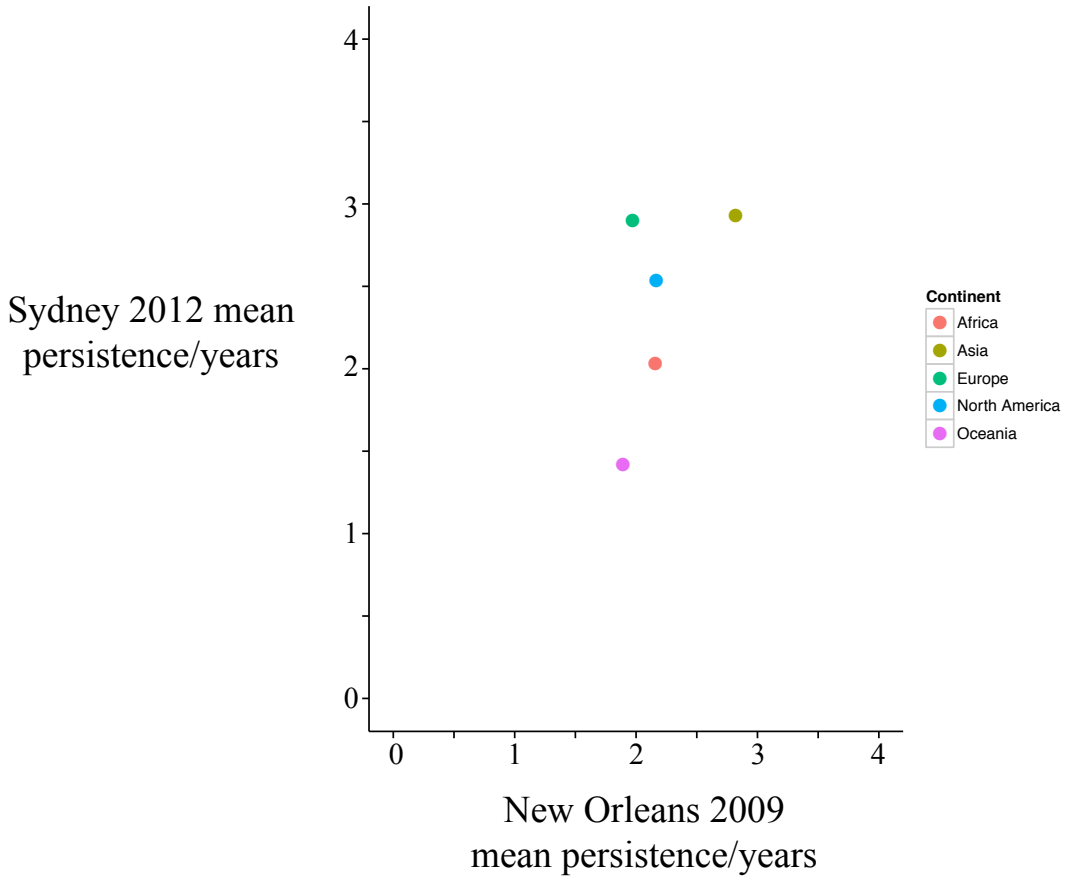


Figure S4.3: No correlation between the average persistence of viral lineages within continents between New Orleans 2009 and Sydney 2012. The mean persistence for New Orleans 2009 lineages within each continent is plotted against the mean persistence for Sydney 2012 viral lineages within the continent.

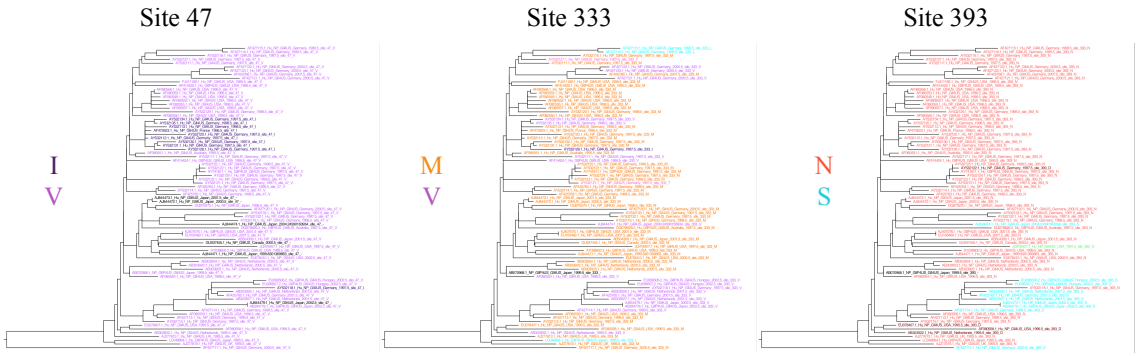


Figure S4.4: Variable sites within US95/96. Coloured trees are shown for each of the sites identified as exhibiting high variability in the US95/96 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.

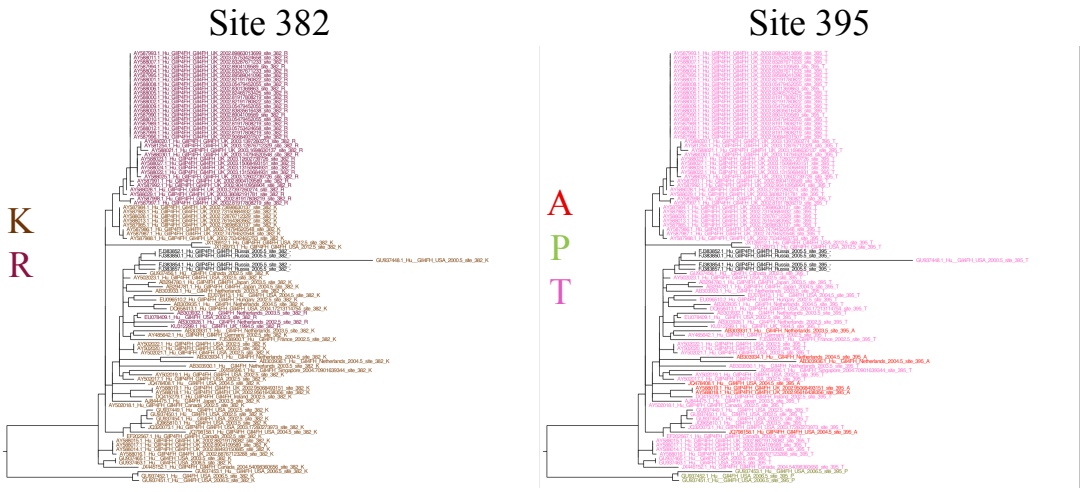


Figure S4.5: Variable sites within Farmington Hills 2002. Coloured trees are shown for each of the sites identified as exhibiting high variability in the Farmington Hills 2002 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.

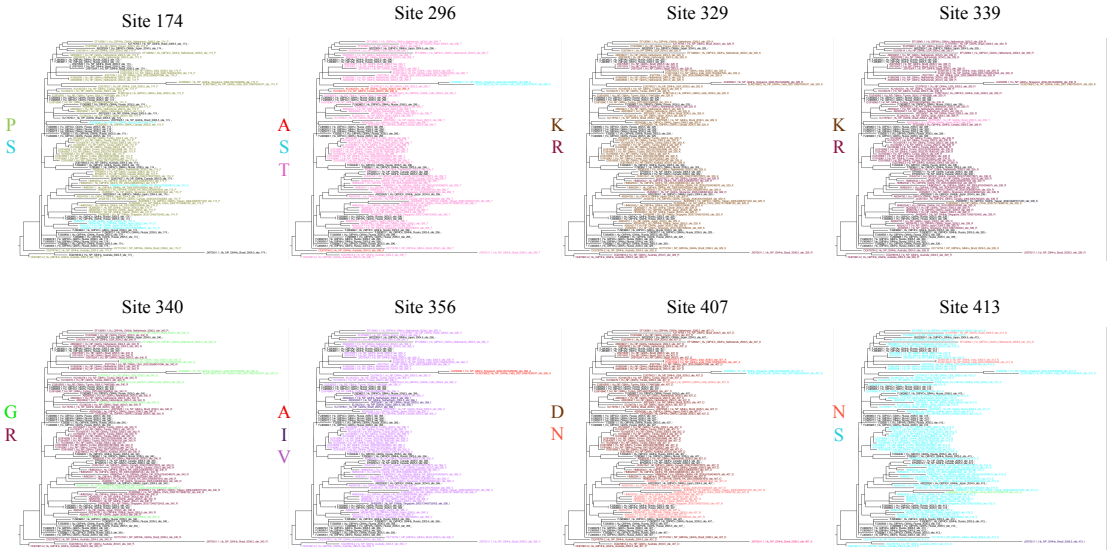


Figure S4.6: Variable sites within Hunter 2004. Coloured trees are shown for each of the sites identified as exhibiting high variability in the Hunter 2004 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.

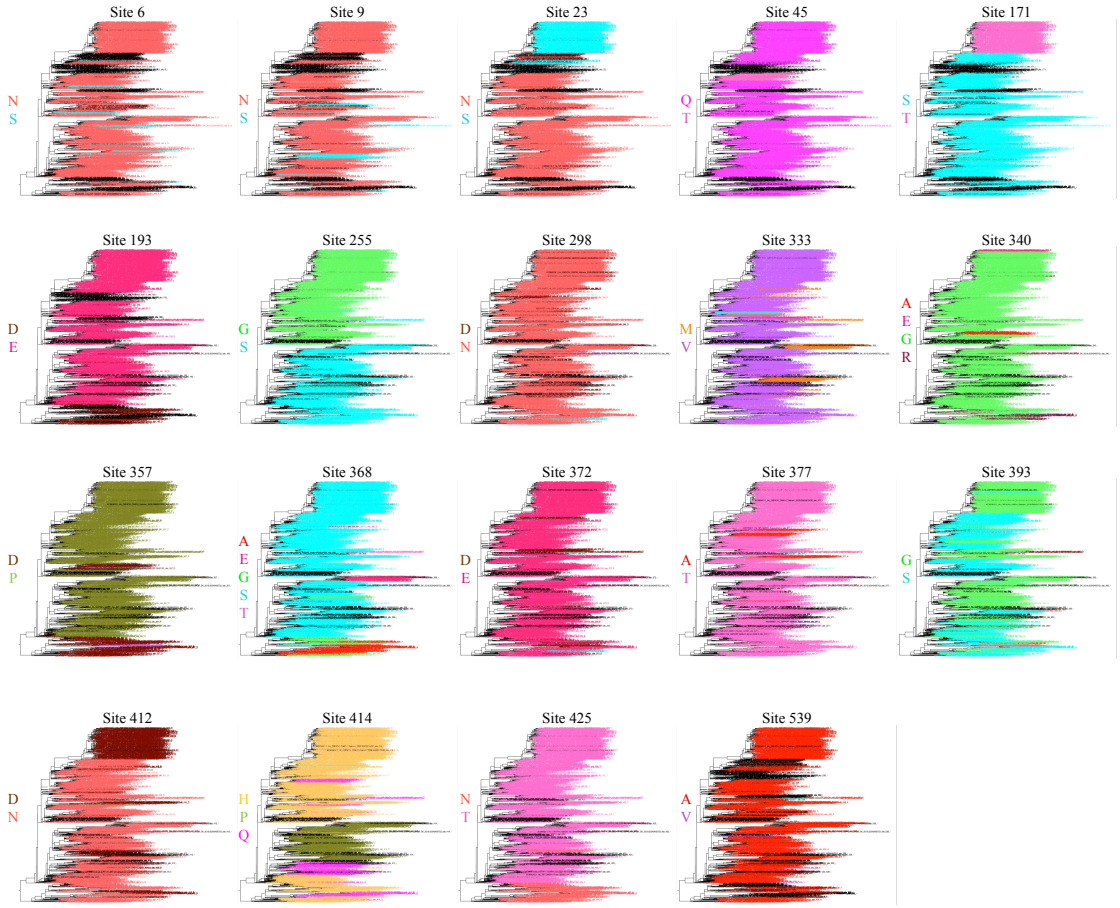


Figure S4.7: Variable sites within Den Haag 2006. Coloured trees are shown for each of the sites identified as exhibiting high variability in the Den Haag 2006 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.

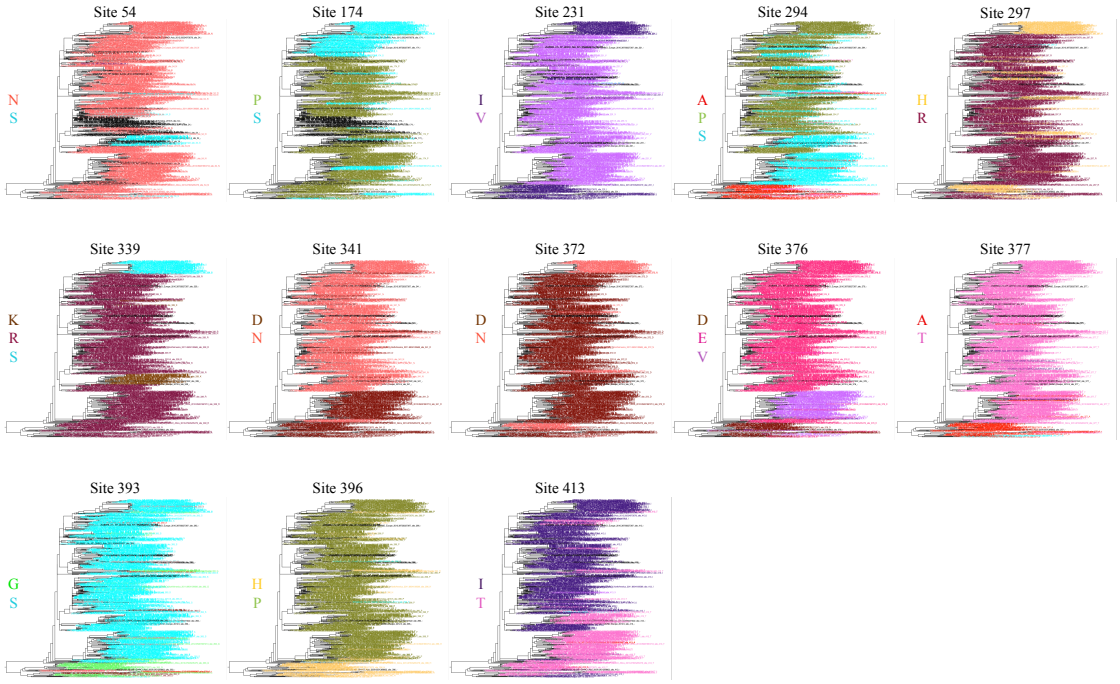


Figure S4.8: Variable sites within New Orleans 2009. Coloured trees are shown for each of the sites identified as exhibiting high variability in the New Orleans 2009 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.

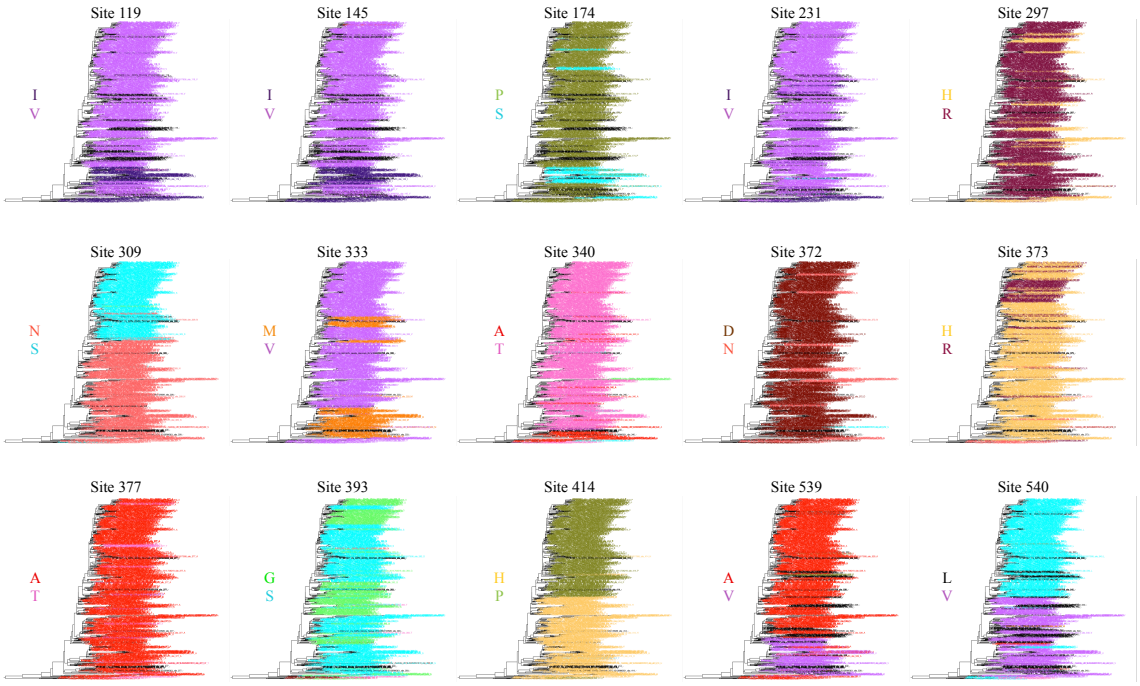


Figure S4.9: Variable sites within Sydney 2012. Coloured trees are shown for each of the sites identified as exhibiting high variability in the Sydney 2012 pandemic GII.4 strain. Each tip is coloured by the amino acid residue in that sequence at the corresponding site, residues are shown next to the tree.